Data and Model Selection to Detect Sparse Events from Multiple Sensor Modalities

Nidhi Parikh[1], Garrison Flynn[1], Dan Archer[2], Tom Karnowski[2], Monica Maceira[2], Omar Marcillo[2], Will Ray[2], Randall Wetherington[2], Michael Willis[2], and Andrew Nicholson[3]

[1]Los Alamos National Laboratory
[2]Oak Ridge National Laboratory
[3]NNSA, Defense Nuclear Nonproliferation R&D

Abstract: Monitoring nuclear facilities is important for nuclear nonproliferation. However, the activities or events of interest are likely to be sparse and occur under variable conditions. In this work, we focus on predicting the power level of a nuclear reactor, where only a few observations are available for the intermediate power levels (10-90%) using data from multiple sensor modalities (seismic, acoustic, thermal, electromagnetic, and effluent). These sensors are positioned near a collocated research nuclear reactor and reprocessing facility at Oak Ridge National Laboratory for the Multi-Informatics for Nuclear Operations Scenarios (MINOS) venture. While combining data from multiple modalities offers opportunities for detecting signals that may not be fully captured by any individual modalities, it also poses a few challenges: 1) Not all of the modalities may provide useful information; 2) A number of features could be computed for each modality and some of these features may be less informative or lack robustness across different reactor startups; 3) Often these reactor startups occur in different environmental conditions which may lead to operation signatures that vary across different startups; 4) Depending upon the physical phenomenon that each sensor is designed to capture, different machine learning models may be most effective. In this paper, we present a systematic approach to select data (including both modality and features) and models to improve prediction of the reactor power level.

Key words: Classification, Data fusion, Persistent Monitoring, Naïve Bayes, Hidden Markov model, Data and model selection

1 Introduction

Monitoring activities at a nuclear facility is important for nuclear nonproliferation. Multi-modal analysis can help identify signals that may not be fully captured by any individual modality. Machine learning models can help relate these signals to facility operations. However, the events of interest are likely to be sparse and occur under variable conditions. In this work, we focus on predicting the power level of a nuclear reactor, where only a few observations are available for the intermediate power levels (10-90%) using data from multiple sensor modalities (seismic, acoustic, thermal, electromagnetic,

and effluent). However, it may be the case that not all of the modalities, or all features extracted from a modality, provide unique and relevant information for the classification problem. A large number of features can often be computed for each modality but some of them may not be informative or lack robustness across different startups. Also, depending upon the physical phenomenon that each sensor is designed to capture, different machine learning models may be most effective. To address this, we present a systematic approach to select both data (including modalities and features) and model to improve reactor power level prediction.

In this paper, focus on predicting the power level of the 85 MW High Flux Isotope Reactor (HFIR) at Oak Ridge National Laboratory. The reactor is pressurized, beryllium-reflected, cooled and moderated by light water [6]. HFIR is a flux-trap type reactor that uses highly enriched ^{235}U as fuel for its 22-26 day cycle [1]. At 2.6 x 10^{15} neutrons per cm^2 per second, it is a world-leading source for steady-state neutron thermal flux [6]. Currently, HFIR is used for production of a wide range of medical radioisotopes [3], material irradiation experiments, neutron scattering, and neutron activation [1]. Four horizontal thermal neutron beam tubes provide neutrons to the neutron scattering instruments [2]. Co-located with HFIR is the Radiochemical Engineering Development Center (REDC). REDC is a multipurpose radiochemical processing and research facility that is home to a variety of laboratory spaces. A key mission supported by REDC is production of radionuclides, including Pu-238 and Cf-252. Target rods are manufactured at REDC and transferred to HFIR for isotope production. After irradiation, target rods return to heavily shielded hot cells at REDC where they undergo a series of dissolution processes to recover isotopes of interest.

2 Multimodal Data

The joint HFIR-REDC complex has been instrumented with a variety of physics-based sensing modalities under the the Multi-Informatics for Nuclear Operations Scenarios (MINOS) project. Sensing modalities include acoustic, electromagnetic (EM), effluent, seismic, and thermal imagery. Each modality persistently measures at a sample rate appropriate for the physical phenomena of interest. Features of the signals, determined with input from subject matter experts, are extracted in time synced thirty-second windows, with a twenty percent rolling window overlap. Every time window is treated then as a sample for classification problem. Measurement details and features extracted are summarized in Table 1. Please refer to [4] for more details on sensors and feature extraction.

3 Methods

3.1 Classification Models

All classifiers used in this work are probabilistic models, and therefore require feature probability distributions conditioned on power level to estimate the power level of the reactor. Gaussian mixture models (GMM) are used to compute conditional feature distributions. GMM is a probability density function that consists of a weighted sum of Gaussian component densities. Please see our previous work [4] for more details about estimating feature distributions using GMMs.

Based on the number of sensors within a modality, sensor bandwidth, and processing method relevant for given the physical phenomena, different number of features are calculated for each modality. For classification purposes, all modalities are weighted equally and within each modality, all features are weighted equally. We evaluated two classifiers: naïve Bayes and hidden Markov model.

Naïve Bayes (NB) is a probabilistic classifier that assumes no temporal dependency among observations (i.e., they are independent of each other). It uses Bayes theorem with naïve assumption about

Table 1. Measurements and features of five disparate sensing modalities.

Modality	Sensor	Feature Summary
Acoustic	3 x IML infrasound sampled at 500 Hz	62 Features • Time signal statistics • Spectral content
EM	PEARSON 411c current monitor ground line sampled at 96 kHz	7 Features • Modulation bands • Fan speeds of two variable speed fans
Effluent	ORTEC Detective 200 High-purity Geranium measuring a 16384 channel gamma spectrum sampled at 1 Hz	15 Features • Isotope count rates
Seismic	GeoSpace tri-axial geophone sampled at 500 Hz	49 Features • Time signal statistics • Spectral content
Thermal	FLIR Ax8 thermal imaging camera sampled at 1/3 Hz	6 Features • Cooling tower basin gradient temperatures • Inlet/outlet pipe temperature differential

the independence of features given the class assignment and assigns the observation to the class that maximizes the probability given the observed feature set.

Hidden Markov model (HMM) [5] is a temporal extension of naïve Bayes classifier. It is used to model discrete-time stochastic process where the state of the process (which is equivalent to a class) is hidden but each state or class generates an observation (i.e., a feature set) which is observable. It make Markov assumption for classes, i.e., the class at time t depends only on the class at time $t-1$. In contrast with NB classifier, HMM takes into account history (a time series of observations) to estimate power level at any time t. As our previous work [4] showed that HMM with 1 minute of history lead to the optimal results, we fix the length of history for HMM to 1 minute for this paper. Please see our previous work [4] for detailed information about the NB and HMM classifiers applied in this study.

3.2 Performance Metric

The multi-class problem posed here is a discretization of a continuous process. Therefore, a distance-based metric for quantifying performance based on the confusion matrix is appropriate. The scoring matrix S is defined as 1 along the diagonal to credit a correct class prediction. Off diagonal elements of S are negative, decreasing to -1, applying an increasingly large penalty based on degree of misclassification distance as shown in Figure 1. Using this scoring matrix we calculate a score for every cycle holdout as follows:

$$Score = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij} x S_{ij}$$

where C is the confusion matrix, n is the number of true classes, and m is the number of predicted classes. Score can range from 1 (all predictions equal truth) to -1 (all predictions are 100% power when truth is 0% power and vice versa). This scoring approach is flexible in the credits and penalties so that a subject matter expert can adjust the score based on appropriate risk for a given scenario.

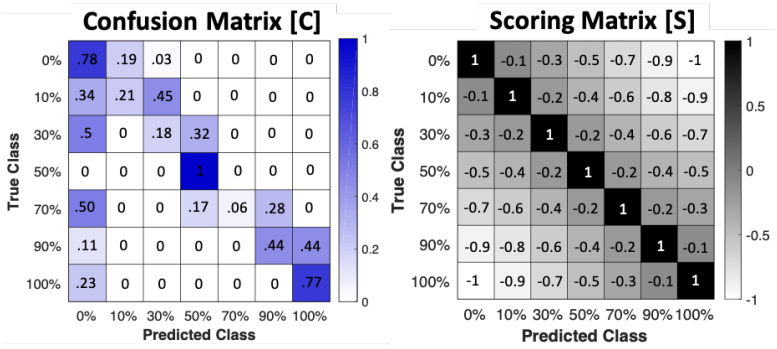

Fig. 1. Confusion and scoring matrices.

3.3 Data and Model Selection

In our previous work [4], we evaluated all possible combinations of modalities using NB classifier and identified top four combinations of modalities that lead to the optimal results. We further evaluated these four modalities combinations using four classifiers (NB, NB sequential, HMM, and an ensemble model that combines the predictions from NB, NB Sequential, and HMM) and showed that irrespective of the models used, a combination of thermal, EM, and effluent modalities leads to the optimal predictions (average performance across cycles). The results also showed that optimal model depends upon the modalities chosen (i.e., HMM led to the optimal performance when only thermal, EM and/or effluent were selected but NB sequential led to the optimal performance when acoustic modality was included). This suggests a two step approach where in the first step, an optimal set of modalities is chosen using the simplest model (i.e., NB classifier) and in the second step, based on the selected modalities, optimal model is chosen.

3.3.1 Interpreting Optimal Modality Combination and Motivation for Data Selection Approach

To understand why thermal, EM, and effluent modality combination leads to optimal performance in our previous work [4], we started by performing correlation analysis. Our hypothesis is that features from these modalities are more informative (highly correlated with power level) and robust (have less variation in correlation with power level across cycles), and these modalities aren't strongly (or at least less so than other modalities) correlated with each other. Thus, contributing a unique piece of useful information for power level prediction.

As Table 2 shows, all of the thermal features are highly correlated with power level (correlation >= 0.7) and also have very low variation in correlation across cycles (i.e., standard deviation of correlation across cycles is less than 0.1). This suggests that it is useful for power level prediction. This is also supported by its rank (17/31; highest ranked individual modality) when using naïve Bayes classifier.

Only a small fraction of seismic features are highly (correlation >= 0.7; 6% of features) or moderately (correlation >= 0.5; 16% of features) correlated with power level (Table 2). This suggests that seismic data may not be very useful to predict power level. This is also supported by its rank (it is ranked the last).

EM is the second most correlated modality with power level. However, only 2 out of 7 EM features have low variation (standard deviation < 0.1) in correlation across cycles (i.e., correlation of most of EM features with power level varies from cycle-to-cycle) so EM by itself is not very useful for prediction (ranked 26/31; 4th individual ranked modality when using naïve Bayes classifier). But most of EM features (26 out of 42 combination of thermal-EM features; 62%) have very low correlation (< 0.1) with thermal. This suggests that it is contributing some unique piece of information that is

Criteria	Number (fraction) of features meeting criteria				
	Thermal	EM	Effluent	Acoustic	Seismic
Number (fraction) of features highly correlated with power level (avg r>=0.7)	6/6 (100%)	3/7 (43%)	3/15 (20%)	14/62 (23%)	3/49 (6%)
Number (fraction) of features at least moderately correlated with power level (avg r>=0.5)	6/6 (100%)	7/7 (100%)	5/15 (33%)	19/62 (31%)	8/49 (16%)
The number (fraction) of features with low standard deviation (<0.1) of correlation with power level across cycles	6/6 (100%)	2/7 (29%)	14/15 (93%)	26/62 (42%)	8/49 (16%)
For features that are at least moderately correlated with power level, the number (fraction) of modality-thermal feature pairs with very low correlation (avg r<0.1)		26/42 (62%)	23/30 (77%)	54/114 (47%)	
For features that are at least moderately correlated with power level, the number (fraction) of modality-EM feature pairs with very low correlation (avg r<0.1)			35/35 (100%)	55/133 (41%)	
NB rank: individual modality	17/31	26/31	18/31	24/31	31/31
NB rank: modality + thermal		3/31	5/31	14/31	21/31
NB rank: modality + thermal + EM			1/31	4/31	10/31

Table 2. Correlation analysis

not captured by thermal. Hence, when combined with thermal, it is quite useful (ranked 3/31; highest ranked two modality combination when using naïve Bayes classifier).

Effluent and acoustic features have about the same level of correlation with power level (i.e., 31-33% of effluent and acoustic features are at least moderately correlated with power level). But most of the effluent features (14 out 15; 93%) have low standard deviation in correlation across different cycles as compared to acoustic features (26 out 62; 42%) and effluent features are less correlated with thermal and EM features as compared to acoustic features (Table 2). This suggests that effluent features may be more useful for prediction as compared to acoustic features. This is also supported by their ranks when combined with thermal and/or EM.

The above analysis helps explain why a particular set of modalities led to the optimal results using correlation analysis. However, the correlation thresholds used for this analysis were chosen somewhat arbitrarily. In this paper, we explore different threshold values for feature and power level correlation to chose the optimal values of these thresholds in combination with modality selection to select both optimal set of modalities and features simultaneously in order to improve prediction of the power level.

Limitations: We used pearson correlation for this analysis which measures linear relationships between variables. There may be informative non-linear relationships between these variables which

are not captured by our analysis.

3.3.2 Data Selection

We use forward selection algorithm for selecting data (both modality and features). High level description of this process is as follows: Initially, no modalities are selected and let's assume average prediction score across cycles when no modalities are selected is 0. Then, iteratively, a modality (along with selected features) is added to the list of selected modalities (and features) unless adding a modality does not improve average prediction score across cycles. The detailed description is below:

At the beginning of iteration 1, as no modalities are selected, we try each modality individually. For each individual modality, we compute average prediction score across cycles using cross validation (CV) with no feature selection (i.e., when all features are included) using NB classifier. Next, we try a number of threshold values for mean and standard deviation (std) of correlation between features and power level to select features and choose the pair of mean and std threshold values that lead to optimal performance in terms of average prediction score across cycles using CV. If feature selection improved results over no feature selection, then we keep track of the optimal threshold values and use them to extract the optimal set of features for the given modality. Once all individual modalities are evaluated, we choose the one that leads to best average prediction score along with the selected features and add it to the list of selected modalities and features.

Suppose, we have data available for five modalities (acoustic, seismic, thermal, EM, and effluent) and iteration 1 selected thermal as the optimal individual modality with four features. Then, in iteration 2, we try to see if adding any other modality to thermal (with the four selected features) can help improve average prediction score. Once again, we try adding a modality both without feature selection and with feature selection (by exploring different threshold values for mean and standard deviation of correlation between features and power level), and select the modality (along with features) that lead to the optimal average prediction score and add it to the list of selected modalities and features if it improves average prediction score across cycles from iteration 1. This process is continued until adding a modality does not improve average prediction score over previous iteration. As we are evaluating addition of a modality to a set of already selected modalities in each iteration, it is not necessary to look at correlation between them. If a new modality does not add new information, adding it will not improve the average prediction score and it will not be selected.

3.3.3 Model Selection

Once optimal modality and features are selected, we use CV to evaluate all models with these modalities and features and select the one that leads to the optimal prediction score across cycles.

3.4 Performance Evaluation

As shown in Figure 2, we use nested CV to select optimal data and model and evaluate the performance of this optimal selection. In n-fold CV, a dataset is divided into n parts (called folds). Each of n folds are heldout in turn for test purpose and the remaining $n-1$ folds are used to train the model which is then evaluated on the heldout fold. The overall performance of the model is measured as average performance across n test folds. As there is temporal correlation within a cycle, each fold consists of a cycle to avoid overfitting. The nested CV here uses two inner CVs. First inner CV is used to select optimal modalities and features and second inner CV is used to select optimal model for the selected modalities and features. Finally, the outer CV is used to evaluate the performance of optimal data and model selected from the inner CV on the heldout cycle.

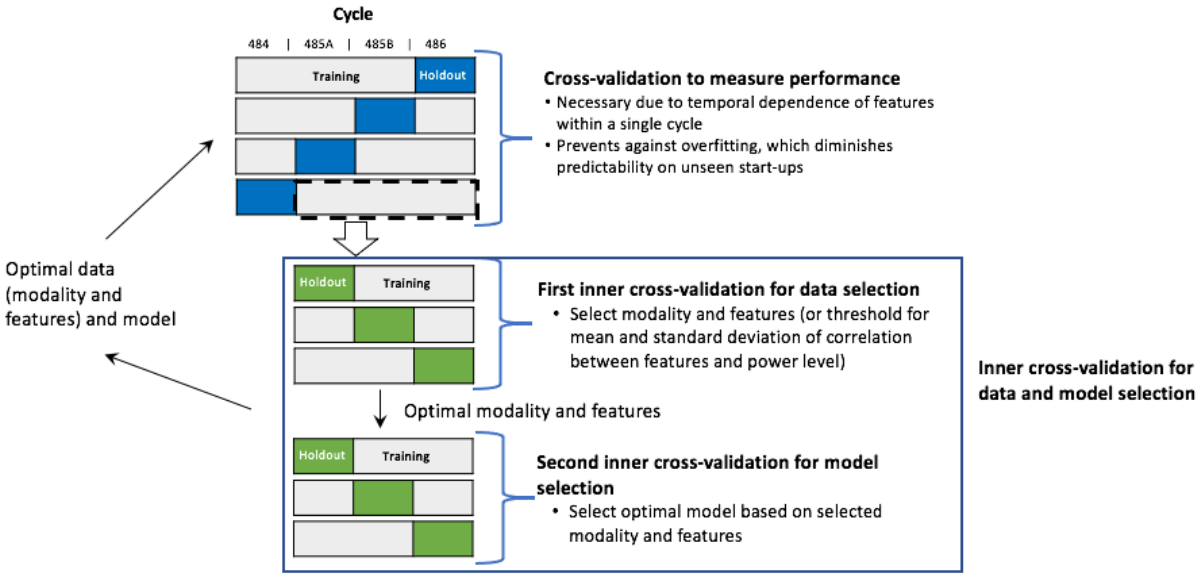

Fig. 2. Nested cross-validation for data and model selection and performance evaluation

4 Results

For each heldout cycle, Table 3 shows the optimal set of modalities selected (along with the order in which they were selected), threshold values of mean and standard deviation of correlation between features and power level used for optimal feature selection, and the optimal model chosen using the training data. It also shows prediction score for the heldout cycle as modalities are selected iteratively using NB classifier and final prediction score with the optimal data (modality and feature) and model. For three out of the four sets of training data (or folds of outer CV), thermal, EM, and effluent were selected as optimal modalities which is consistent with our previous results [4] which found the same modalities as the optimal set followed by thermal, EM, effluent, and acoustic which were selected as optimal modalities for the remaining set of training data (fold of outer CV). The results also show that as each modality is added to the set of selected modalities, it significantly improves prediction score when using NB classifier.

Simultaneous with modality selection, optimal features for individual modalities are selected. The frequency of features selected is shown in Table 4. For thermal, 50% of features are selected every cycle, 83% at least two cycles, and 100% at least one cycle. Effluent features are downselected to include 33% of features every cycle, and 100% of features at least two cycles. EM selects 43% of features every cycle, 86% at least three cycles, and 100% at least two cycles. Acoustic has the most significant downselection, eliminating all frequency domain features and selecting only time series statistics from the three sensors for 18% of features to be selected, but only for one cycle.

Table 3 shows that three times out of four, NB is selected as the optimal model, and HMM is selected once. In our previous results [4] without feature selection, HMM was often selected as the optimal model when thermal, EM, and effluent modalities were used. These new results suggest that with careful feature selection, NB (a simpler model) could outperform HMM.

Table 3. Optimal Modalities, Model, and Performance

Heldout cycle	Selected modalities in order of their selection	Mean Correlation threshold	Std. Correlation threshold	Score with optimal data and NB model	Optimal model	Score with optimal data and model
484	Thermal	None	None	-0.0702	NB	0.3609
	Thermal, Effluent	None	None	0.1377		
	Thermal, Effluent, EM	None	None	0.3609		
485A	Effluent	0.6	0.1	0.0696	HMM	0.5365
	Effluent, Thermal	0.9	0.1	0.5654		
	Effluent, Thermal, EM	0.5	0.3	0.6441		
485B	Thermal	0.8	0.2	0.3616	NB	0.5815
	Thermal, Effluent	None	None	0.4908		
	Thermal, Effluent, EM	0.6	0.3	0.5815		
486	Thermal	0.9	0.1	0.0215	NB	0.4318
	Thermal, Effluent	0.6	0.2	0.3610		
	Thermal, Effluent, EM	None	None	0.3651		
	Thermal, Effluent, EM, Acoustic	0.9	0.1	0.4318		

Table 4. Frequency of features selected.

Modality	Number of Cycles Feature is Selected			
	4	3	2	1
Thermal	Basic C, Basin Average, Inlet-Outlet Differential	-	Basic B, Basin D	Basin A
Effluent	41Ar, 135mXe, 137Xe, 138Cs, 138Xe	- -	88Kr, 90Kr, 131I, 132I, 133I, 134I, 135I, 135Xe, 139Ba, 139Xe	
EM	D, Fan C Speed, Fan D Speed	B, C, E	A	
Acoustic	-	-	-	Time series features: RMS, peak velocity, 90th and 95th percentiles

Figure 3 shows ground truth and predicted power level using optimal data and model for all four cycles. The results show that optimal data and model selection leads to accurate power level predictions for cycles 485A, 485B, and 486 (although with a few exceptions). For cycle 486, the predicted power varies rapidly between 0, 10, and 30% when the reactor is at 10% power hold and between 0, 90, and 100% when it is at 100% power. This is because the selected optimal model NB does not take into account temporal dependence among observations. In our previous work, we have shown that by taking into temporal dependency, HMM was able to help reduce this problem. This suggests that the selected model (NB) is not actually optimal for cycle 486.

The results also show that the optimal data and model does not lead to correct predictions consistently when the power level is 70% for cycles 485A, 485B, and 486. As explained in our previous work [4], this is because of a disproportionally large power hold at 70% for cycle 484 as compared to other cycles and a number of process being tested during cycle 484 (which does not happen usually), making cycle 484 quite different from the other cycles. This is also the reason why predictions for cycle 484 aren't accurate. We believe that having more cycles and/or removing cycle 484 from this small set of training data could potentially help solve these problems.

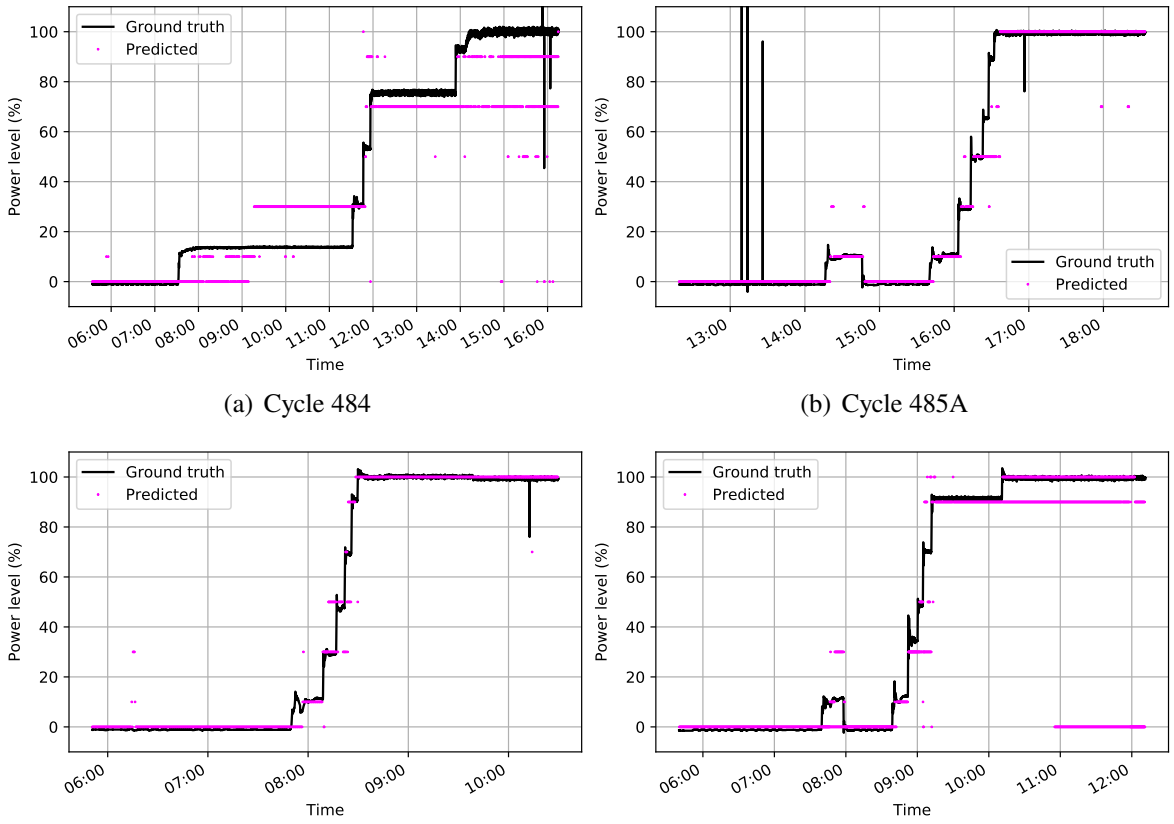

Fig. 3. Ground truth and predicted power level with optimal data and model. Notes: 1) The predicted power levels are shifted up slightly so that they are not overlayed on top of the ground truth and can be differentiated easily. 2) X-axis is on different scales for subfigures (a), (b), (c), and (d).

5 Conclusions

Multiple modalities and machine learning models could help monitor nuclear facilities for non-proliferation. However, not all modalities (and features) may be informative and depending upon the physical phenomenon that each sensor is designed to capture, different models may be most effective. In this work, we focused on predicting the power level of a nuclear reactor using multiple modalities and presented a systematic approach for selecting both data (modalities and features) and model to improve prediction. The results show (and match our previous results) that thermal, EM, and effluent modalities lead to the optimal performance followed by thermal, EM, effluent, and acoustic. The results also show that by careful feature selection NB model can outperform HMM (an extension of NB that takes into account temporal dependence of observations). The methodology presented allows for the optimal model to be determined given all observations available, and has the advantage of easily updating model optimization as new training sets becomes available. Finally, the approach developed provides a mean of prioritizing sensing modalities and features for future collection campaigns.

Acknowledgements

This work was performed under the NNSA Office of Defense Nuclear Nonproliferation R&D Multi-Informatics for Nuclear Operations Scenarios (MINOS) Venture.

References

[1] N. S. Directorate. High flux isotope reactor. https://neutrons.ornl.gov/hfir. Accessed: 2019-05-15.

[2] W. T. Heller, V. S. Urban, G. W. Lynn, K. L. Weiss, H. M. O'Neill, S. V. Pingali, S. Qian, K. C. Littrell, Y. B. Melnichenko, M. V. Buchanan, et al. The bio-sans instrument at the high flux isotope reactor of oak ridge national laboratory. *Journal of Applied Crystallography*, 47(4):1238–1246, 2014.

[3] F. R. Knapp Jr, S. Mirzadeh, A. Beets, and M. Du. Production of therapeutic radioisotopes in the ornl high flux isotope reactor (hfir) for applications in nuclear medicine, oncologyand interventional cardiology. *Journal of radioanalytical and nuclear chemistry*, 263(2):503–509, 2005.

[4] N. Parikh, G. Flynn, E. Casleton, D. Archer, J. Johnson, T. Karnowski, A. Nicholson, M. Maceira, O. Marcillo, K. Myers, W. Ray, R. Wetherington, and M. Willis2. Predicting the power level of a nuclear reactor using atime series-based approach. *Institute of Nuclear Materials Management*, 2020.

[5] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *PROCEEDINGS OF THE IEEE*, pages 257–286, 1989.

[6] N. Xoubi and R. Primm III. Modeling of the high flux isotope reactor cycle 400. *ORNL/TM-2004/251, Oak Ridge National Laboratory*, 2005.