Unsupervised Causal Deep Learning-Based Anomaly Detection in Nuclear Power Plant Applications

Abhinav Saxena[1], Helena Goldfarb[1] and Jeffrey Clark[2]

[1]*GE Research, Niskayuna, NY, 12309, USA*
asaxena@ge.com
goldfarb@ge.com

[2]*Constellation Inc., Baltimore, MD 21231, USA*
Jeffrey.Clark2@constellation.com

ABSTRACT

Nuclear power generation will be key to meeting carbon free energy transition goals. However, nuclear power must provide agility and flexibility to fluctuating power demands when other sources of carbon-free energy like solar and wind may not be as flexible. This has led to small modular reactor (SMR) development where nuclear power will be generated from a distributed fleet of smaller reactors where units can be brought online or offline as needed. Due to its high operational and maintenance (O&M) costs as it is, a distributed fleet will put additional cost burden if remote monitoring and crew sharing is not enabled. This requires prognostics and health management (PHM) capabilities such as early warning, diagnostics, and prognostics to enable predictive maintenance with high accuracy. Typically, monitoring solutions are developed on component and subsystem levels targeting specific failure modes. However, it is argued that a systemwide monitoring, in addition to specific targeted analytics, would be of key importance. This paper presents a deep-causal unsupervised anomaly detector that has been successfully applied in various aerospace and renewable energy applications. In this paper we share our experience applying this method on a nuclear power plant (NPP) application. Specifically, we share how we dealt with practical challenges of data quality, ground truth labeling, performance evaluation and field validation in an unknown-unknown setting where prior knowledge of failures and failure modes were not available to begin with.

1. INTRODUCTION

Nuclear power is considered a carbon-free energy source, yet it has one of the highest levelized cost of electricity (LCOE) compared to other generation sources (NEI, 2021). Regulatory requirements originating from safety and security concerns require an order of magnitude higher levels of staffing per plant as compared to their fossil fuel counterparts. Alignment with decarbonization-based climate goals is projected to significantly rely on nuclear power as part of the energy mix. This, however, will only be feasible if nuclear generation's O&M costs can be substantially reduced without impacting safety risks.

1.1. PHM for Nuclear Industry

Traditionally large-scale nuclear power plants are designed for baseload operation, meaning they run at a constant power output and are not easily adjusted to follow load demand variations. SMRs and microreactors are a type of nuclear power technology that offers a modular and more flexible approach to generating electricity compared to traditional large-scale nuclear reactors. They are designed to be compact, transportable, and scalable, with a power output typically ranging from tens to a few hundred megawatts. Smaller form factor reactors can be deployed individually or in clusters, depending on the electricity demand of a particular location. This scalability allows for more flexibility in matching power generation capacity with varying needs, making them suitable for a wider range of applications, including remote communities, industrial sites, and smaller grids. This concept, however, will not scale as staffing levels cannot simply be scaled down proportionally from a large plant, thereby resulting in diseconomy of scale and inflated cost burden. Remote monitoring with centralized crew sharing is seen as a key enabler for this concept. Remote monitoring and early warning from a PHM system would be essential to enable condition-based maintenance and outage planning. However, as with any critical application, uncertainty quantification, explainability, and runtime prediction robustness are key challenges to field deployment.

1.2. Literature Review

Remote condition monitoring has been pursued in the nuclear industry and more generally in power plants for a number of years. Most of these approaches have used traditional machine learning-based classification and regression models that do not scale well in real industrial environments with required levels of prediction accuracy (Yadav, et al., 2021) and are built to be static and fixed input based. Model-based approaches work well when models are developed for specific components and failure modes. Generalized multivariate anomaly detection techniques providing broad coverage of components and failure modes are not widely discussed in nuclear contexts.

For instance, a simple first order polynomial regression model was developed as an anomaly detector using data with multiple failure modes (Moleda, Momot, & Mrozek, 2020) on a boiler feed water pump in a coal plant. Their approach required adjusting thresholds for each failure mode and a clear assessment of false detection and missed events was not provided. In (Ramuhalli, Walker, Agarwal, & Lybeck, 2021) a number of multivariate time series deep learning models are discussed for prognostics. Their findings indicate gap in systematically learning optimal model hyperparameters that would generalize over a wider set of operational conditions expected to be seen over time.

As discussed in section 1.4, the intent behind this work was to apply a modeling framework successfully deployed in other industry domains to nuclear context without requiring new development work. Here we document our experience in terms of ease of model training and performance assessment from real operational environment towards establishing feasibility of deployment in nuclear plants.

1.3. Opportunity

Ongoing research and development in the field of advanced nuclear reactor designs, including small modular reactors, aims to enhance operational flexibility to integrate better with the grid of the future. Capability of a plant to adjust its power output within safety envelope as demand for electricity changes is key to such flexibility. The SMR designs incorporate inherent safety features and flexible operation modes, making them better suited for flexible grids with distributed generation sources such as wind and solar. GE Hitachi is currently developing BWRX-300, one of the leading SMR designs, which is expected to become operational by 2028 (GE Hitachi, 2017). Since the reactor is in its design phase, this presents a unique opportunity to develop and integrate PHM from an early phase. This reactor will be a first of its kind and there is no operational experience or reliability data available. Furthermore, PHM technologies even if proven to be successful in other industrial domains must be evaluated and adapted to a nuclear plant domain. Under DOE ARPAE's funded program called Generating Electricity Managed from Intelligent Nuclear Assets (GEMINA) a number of machine learning (ML) and artificial intelligence (AI) techniques have been developed and evaluated towards enabling predictive maintenance using BWRX-300 as a reference design. Operational data from existing Boiling Water Reactor (BWR) plant fleets as well as plant simulators were collected for PHM analytics development. This paper describes results from one of the many workstreams that focused on unsupervised systemwide monitoring using historical plant data.

1.4. Research Questions Investigated

Industrial systems such as power plants generate large amounts of data that grow in size over several decades of operation. Throughout the life cycle of a plant numerous events and activities take place that affect measurement data. Labeling these multivariate data for machine learning tasks is next to impossible rendering the supervised class of methods of little use. In this work our intention was to use unsupervised methods and evaluate whether the outcomes can be meaningfully integrated into existing maintenance workflows. Specifically, the following research questions were pursued:

- whether machine learning based health monitoring models can be demonstrated to generate an early warning in examples where faults went undetected until a late stage
- whether generation loss from forced outage could be avoided/minimized in economically meaningful way
- whether we can limit the false positive rate such that desired economic benefit can be achieved as set forth by performance targets based on cost-impact analysis of the historical maintenance burden data.

We present our approach and results next. To the extent possible, we have collected and analyzed plant history to generate relevant context including discussion of these results with plant subject matter experts (SMEs) towards validation and further scope of expansion/refinement of the algorithms.

It must be noted that to protect plant identity and company confidential information, data identifiers have been omitted and data are anonymized on purpose.

2. PROBLEM FORMULATION

The initial phase of the project engaged in analyzing large amounts of historical events data in existing nuclear plants as reported in Institute of Nuclear Power Operations (INPO) reports (INPO, 2021). Events resulting in generation loss (due to forced shutdown or power reductions) were identified to determine sites and timeframes of interest to obtain plant operational data.

2.1. Field Event Description

In one particular scenario, an anomaly was detected during a routine tour and the site staff had to scramble to repair the near failed turbine driven reactor feedwater pump. Inspection of the inactive thrust bearing revealed non-uniform wear of pin and leveling pads and the evidence pointed to a lightly loaded thrust bearing. This was a *Maintenance Rule Functional Failure* (10 CFR 50.65 (NRC), 1974) that involved exceeding a plant-level monitoring criterion and was determined to be a consequential equipment failure even though there were no industrial safety or radiological consequences as a result of this event.

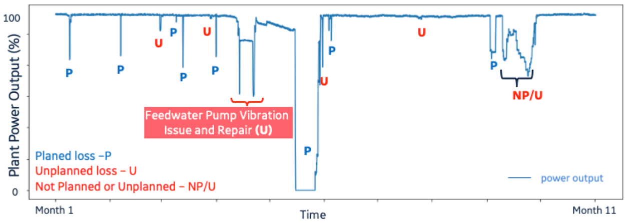

Figure 1. Operational history for a 10-month period depicted through normalized plant power output.

As shown in Figure 1, the event of interest (feedwater pump vibration related failure) is marked as one of the long unplanned outages (U). It must be noted that a planned outage was scheduled within a month of this event. Hence, the key value proposition of a health monitoring would be to detect impending failure early enough to prevent pump damage from exceeding plant level monitoring criterion and perhaps continue operating with that pump until the planned outage period. Failures of these kinds are known to show exponential growth rates as degradation progresses, hence even a few days of early warning would be sufficient to catch this degradation earlier (than when it became noticeable to the area operator) and identify a potential intervention to allow plant operation to continue without violating maintenance criterion.

2.2. Data Description

Industrial plant data evolve over time due to usage, wear and tear, and maintenance and repair operations. This requires any monitoring model to continuously update and adapt as plant characteristics change. In this work we assume that plant characteristics do not change significantly enough in less than a year timescale and any model trained or updated within the last six to eight months can be considered current unless there are specific maintenance or upgrade activities carried out. Using our reference failure event described above, we collected operational data from about six months prior to the event and also few months following the repair. After consulting with subject matter experts from the plant, we concluded that examining the operational history spanning 6 to 8 months would likely encompass the majority of operational deviations. This timeframe also allows us to work within the limitations of data volume. It's important to note that while collaborating in a research endeavor across different organizations, the data access constraints we encountered were more aligned with the overall program goals rather than being purely technically oriented.

It was previously shown in other industrial domains (Huang & Kasiviswanathan, Streaming anomaly detection using randomized matrix sketching, 2015), periodic model updates have proven effective in maintaining low false positive rates while not missing any key events of interest. Therefore, we make an assumption that similar model management techniques will be applicable in nuclear plant context and that our model needs to be trained/updated with data from recent timeframes.

Plant data are typically gathered at various time resolutions depending operational needs. However, historical data are often compressed (in time or frequency domains) for long term archiving. In this scenario, due to various technical and resource availability constraints, we could obtain data only at an hourly resolution for the requested period. While this resolution is not ideal for detecting high-frequency and fast progressing failures, our hypothesis was that mechanical failure of this kind often exhibit symptoms over several days before catastrophic failure and we may be able to catch it soon enough even with this resolution. Specifically, we investigated the pump failure issue with a *temporal deep causal anomaly detection* method described in the next section. A total of 35 channels of time-series data were made available for the requested 10-month period. Additionally, periods of planned and unplanned outages were collated by identifying issue reports (IRs) generated from this time range for generating additional context during performance validation. A number of statistical data analyses were carried out to detect correlated parameters, assess noise levels, or ascertain data integrity and data quality. We omit details of data-preprocessing steps (these data cleaning steps are typical to all data-science projects) to focus on health monitoring model development and analysis in this paper.

3. TECHNICAL CHALLENGES

The chosen scenario for constructing the model was solely based on an incident related to vibrations in the feedwater pump (FWP). However, as depicted in Figure 1, there were numerous planned and unplanned power de-rates and shutdowns during the span of 10 months, which is a common occurrence in nuclear power plants. While unsupervised anomaly detection methods are effective at identifying previously unseen irregularities, in their early stages, they often generate numerous alerts that lack practical significance or operational interest. Examples of such alerts include deliberate power reductions, routine chemical dosing, or planned maintenance activities conducted online.

Such alerts require domain-driven post processing to isolate true anomalies from those of less significance (often considered as false positives). At this stage the key bottleneck is domain validation that requires a lot of effort from SMEs combing through documentation of plant history. However, as more examples of a class of anomalies accumulate over time through these types of validation activities, a supervised approach to detection and diagnosis can be effective in improving the accuracy and explainability of a model.

4. TECHNICAL APPROACH

4.1. Unsupervised Anomaly Detection

While we are working with a specific failure mode, our goal is to build a generic anomaly detector for the pump system without explicit knowledge of a failure mode. Generally, it is common industry practice to build monitoring solutions for known failure types, however, in autonomous paradigms the system must be guarded against all anomalies so appropriate attention can be given in timely manner. Traditional approaches to building predictive maintenance models have focused on addressing known failure modes on key components and are limited to low dimensions (few sensor inputs). They fail to scale when a larger number of subsystems or components in a plant require coverage against a larger set of respective failure modes. It must be noted that for effective remote monitoring coverage should be provided against all plant systems to effectively reduce labor headcount. Furthermore, upwards of several hundreds of measured or derived parameters are available from a typical reactor subsystem. At the cutting-edge of deep learning for time series data, GE Research developed deep causal models that encode hi-dimensional (several hundred variables) multivariate causality between model input parameters to detect anomalies (previously seen or unseen). These twins have been successfully demonstrated in other safety-critical domains like aviation jet engines and wind turbines (Huang & Kasiviswanathan, Streaming anomaly detection using randomized matrix sketching, 2015) (Huang, Yoo, Yu, & Qin, 2015). Unlike many other available deep learning architectures, this architecture has been designed to encode time-series characteristics that are often important for health assessment in industrial settings (Huang, Yan, Wang, & Xue, 2018)For instance, all measurements (e.g., currents, temperatures, pressures, flows from reactor core, etc.) can be used together to generate a causal graph of the nominal core operations. In the event of anomalies, relationships between related measurements are likely to change, which reflects as changes in causal graph and identify the parameters that have deviated from their nominal relationship with other parameters. Identified parameters provide explainability necessary to drive diagnosis and corresponding corrective actions.

4.2. Deep Causal Anomaly Detection Model Architecture

Full technical and implementation details of the model are provided in (Huang, Yan, Wang, & Xue, 2018) and (Feng Xue, 2020), we explain the key intuition behind this architecture in this paper. This model learns variable association by modeling a nonlinear Granger causal graph, where each node represents a variable in the time series, and each edge is directed and weighted, describing the Granger causality (Granger, 1969) between the two connected nodes. Figure 2 illustrates the model architecture that consists of three modules.

Module 1 learns temporal features by implementing a residual neural network (ResNet) (Kaiming He, 2016), (Alex Tank, 2018) to learn univariate non-linearity. Module 2 learns a Granger causal graph form "*normal"* timeseries, i.e. determines all contributing temporal features (output from Module 1) that contribute to predicting values of a given parameter and does so for all parameters parallelly. Finally, Module 3 implements a time-series regression, which combines the output of Module 2 in a non-linear combination to predict *n* future values of a parameter.

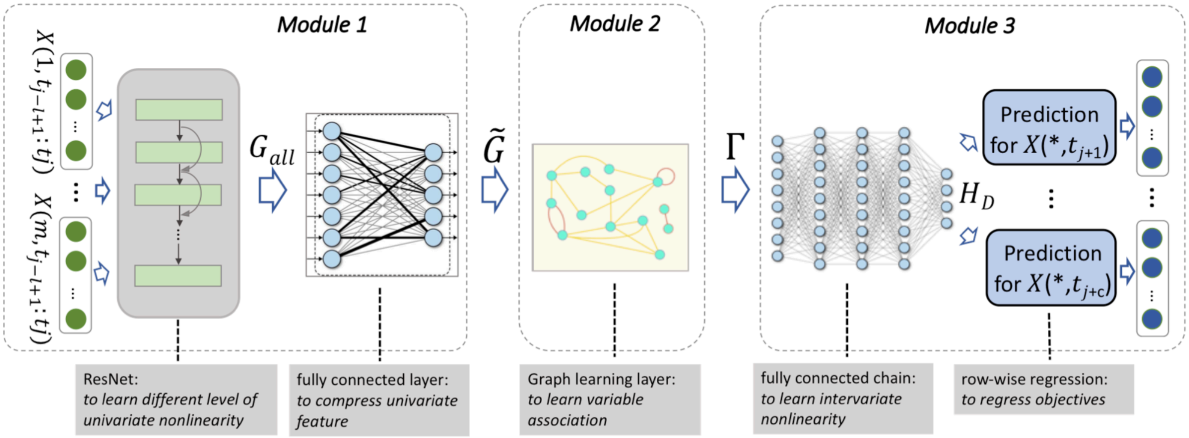
Module 1
Module 2
Module 3
$X(1, t_{j-l+1}:tj)$
$X(m, t_{j-l+1}:tj)$
G_{all}
$\tilde{G}$
Γ
H_D
Prediction for $X(*,t_{j+1})$
Prediction for $X(*,t_{j+c})$
ResNet: to learn different level of univariate nonlinearity
fully connected layer: to compress univariate feature
Graph learning layer: to learn variable association
fully connected chain: to learn intervariate nonlinearity
row-wise regression: to regress objectives

Figure 2. Deep causal anomaly detector network architecture (Huang, Yan, Wang, & Xue, 2018).

Once trained, the network is used as anomaly detector for any incoming input by predicting *n* future values of all parameters and computing residuals from actual values for *n* future time instants (note this results in a slight delay of *n*-timesteps in assessing an anomaly). Furthermore, analyzing the parameters with highest residuals and their relative contributions to the anomaly score we determine presence of an anomaly and the signature of that anomaly. This process is illustrated with the feedwater pump data in the next section.

5. APPLICATION & RESULTS

The deep causal anomaly detection neural network described above underwent training using data extracted from segments where no identified issues were present. In a standard operational context, achieving this involves employing an extended historical time-series record of the asset, assuming that a substantial portion of this extended history accurately represents normal operational behavior. However, given the limitations in available historical data in this instance, we isolated periods characterized by normal behavior for training. These selected periods were intentionally chosen to be distant from repair and outage events.

The complete collection of normal data was subsequently divided into training and validation sets, utilizing an 80-20 split. The remaining data was never shown to the model and was included in test set for evaluation. The network's training was geared towards predicting parameter values for the upcoming two hours, relying on a moving window encompassing the previous five hours. The optimization process aimed to minimize the Mean Squared Error (MSE) as the loss function, continuing until a predefined training termination condition was met for both the training and validation loss.

Following successful training, the model was employed to assess the entirety of the 10-month time-series. This evaluation involved generating prediction residuals for all output parameters. Anomalies were identified based on two primary criteria:

1. An aggregate of residuals surpassing a predetermined threshold.
2. A significant departure in the distribution of residuals from uniformity across all residuals.

Aggregated residuals that fulfilled both anomaly criteria 1 and 2 were denoted as alerts, visually represented as red dots in the top graph of Figure 3. These alerts were further classified in four classes of alerts based on alert signatures. Instances where only one or no criterion was satisfied did not trigger alerts, and were respectively indicated by blue and gray dots.

The lower chart in Figure 3 illustrates the plant's power output, providing a reference point and contextualizing the plant's historical performance, including planned or unplanned outages, de-rates, repairs, and more, thus providing context during the analysis of results. The subsequent step encompassed the analysis and summarization of the alerts generated and is presented next.

5.1. Alerts Analysis

Table 1 shows an aggregated output generated by the model. Each row represents an alert or a group of alerts persisting over more than an hour period (called anomaly period), for which anomaly criteria were met. Column 1 shows average anomaly score for all alerts in an anomaly period. Parameter 1 identifies the input that contributes to the anomaly score the most (e.g. in row 1, `FEEDWATER DISSOLVED O2` parameter contributes to 84% of the total anomaly score of 0.33). Based on past experience, the top 3 contributing parameters and their contribution percentages are generally enough in identifying the nature of the anomaly. The last two columns are the output of the validation exercise, where we sought plant SME input in understanding the cause of the anomaly and if that matched with known events in the plant history, and subsequently mark the anomaly as True Positive, False Positive, or a Positive (late detection) for performance quantification.

As color coded alike in Table 1 and Figure 3 below, a total of 13 anomaly periods spanning 4 types of anomalies were identified. Apart from four alerts due to elevated oxygen levels, all others were deemed relevant alerts in which plant's systems engineers would be interested from predictive maintenance perspectives. Although, our model was clearly able to catch elevated dissolved oxygen on specific days, it was determined to be an effect of operator intervention to manually adjust oxygen level. Analytically our model detected unusual changes (in oxygen level) correctly, it failed to detect the deteriorating oxygen level with early warning and only alerted after intervention was applied, we treat these as late detections. Following this observation, we have agreed to expand the model to include additional chemistry parameters if such data are available in the database. Otherwise, we will improve our model to avoid alerts on this condition, as these are not directly pump related issues. SMEs noted that incorrect oxygen levels are known to lead to cavitation issues on pump internal parts, but whether any correlation exists here needs separate investigation. Quantification of model performance is provided next.

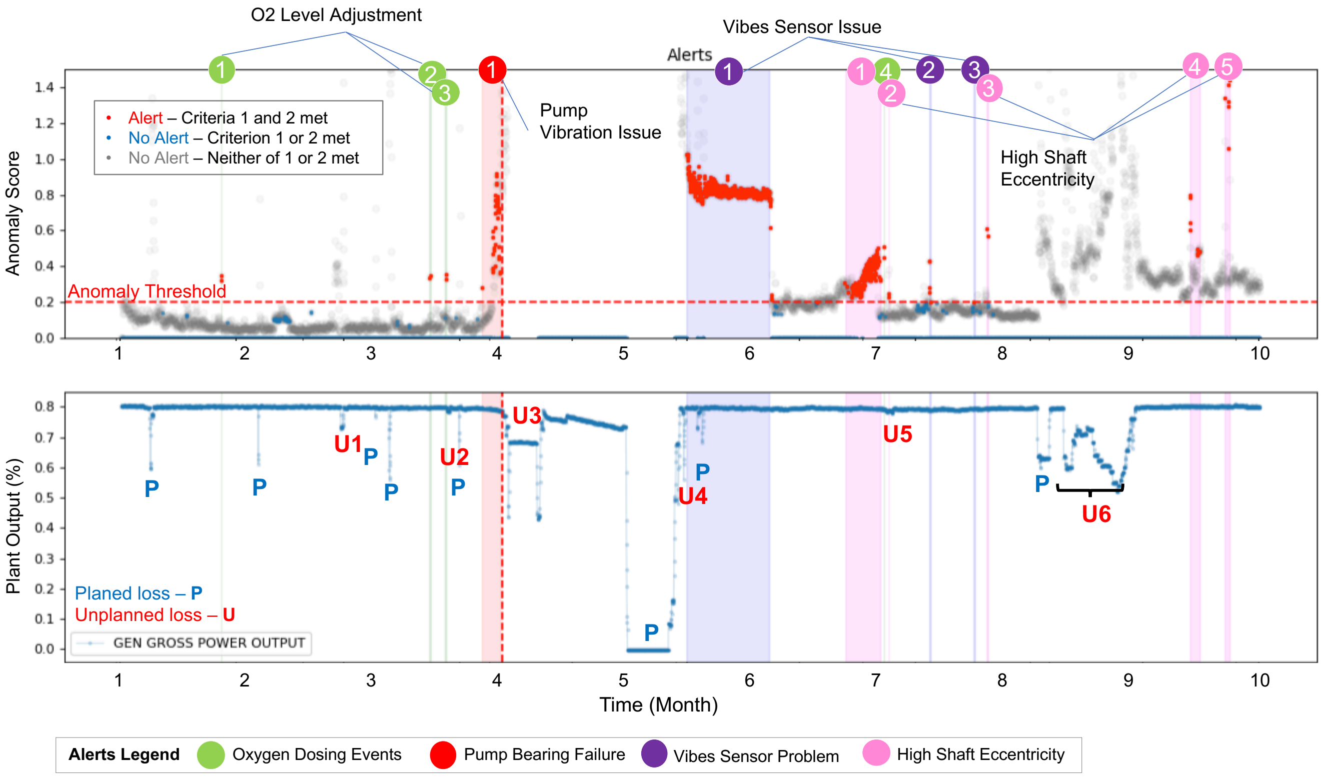

Figure 3. Summary of alerts generated by FWP model over a 10-month period from an existing NPP.

Table 1. Summary of alerts, dispositions and SME feedback for FWP model output.

Avg Anomaly Score	Parameter1	Top parameter contribution	Parameter 2	Parameter 3	Params 2+3 contribution	SME Notes	Resolution
0.33	FEEDWATER DISSOLVED 02	84%	RFP 1B INBD SHAFT VIB	TDFWPBVP	8%	oxygen level was adjusted manually	positive
0.34	FEEDWATER DISSOLVED 02	76%	TDFWPBVP	RFP 1B INBD SHAFT VIB	12%	oxygen level was adjusted manually	positive
0.34	FEEDWATER DISSOLVED 02	82%	RFP 1B INBD SHAFT VIB	TDFWPBVP	16%	oxygen level was adjusted manually	positive
0.58	RFPT 1B ECCENTRICITY	71%	TDFWPBVP	RFP 1B INBD SHAFT VIB	13%	feedwater pump issue (4/13)	true positive
0.84	RFP 1B INBD SHAFT VIB	74%	RFP 1B OUTBD SHAFT VIB	TDFWPBVP	15%	RFP 1B INBD SHAFT VIB sensor is broken, reporting zero vibration.	true positive
0.29	RFPT 1B ECCENTRICITY	56%	RFP 1B OUTBD SHAFT VIB	RFP 1B INBD SHAFT VIB	32%	Pump was wobbling a little bit resulting in eccentricity higher than normal. The issue was fixed after rebalancing.	true positive
0.48	FEEDWATER DISSOLVED 02	59%	RFP 1B OUTBD SHAFT VIB	RFP 1B SEAL WTR FLOW	18%	oxygen level was adjusted manually	positive
0.24	RFPT 1B ECCENTRICITY	53%	RFP 1B OUTBD SHAFT VIB	RFP 1B INBD SHAFT VIB	30%	Pump was wobbling a little bit resulting in eccentricity higher than normal. The issue was fixed after rebalancing.	true positive
0.32	RFP 1B OUTBD SHAFT VIB	75%	RFPT 1B HP STOP VLV TEMP	RFP 1B INBD SHAFT VIB	9%	Outboard vibration sensor power supply out	true positive
0.20	RFP 1B OUTBD SHAFT VIB	50%	TDFWPBVP	RFPT 1B ECCENTRICITY	30%	Outboard vibration sensor power supply out	true positive
0.59	RFPT 1B ECCENTRICITY	77%	RFP 1B OUTBD SHAFT VIB	TDFWPBVP	16%	Increased eccentricity observed	true positive
0.56	RFPT 1B ECCENTRICITY	58%	RFP 1B OUTBD SHAFT VIB	RFPT 1B HP BRG SHAFT VIB	27%	Increased eccentricity observed	true positive
1.52	RFPT 1B ECCENTRICITY	79%	FEEDWATER DISSOLVED 02	RFPT 1B HP BRG SHAFT VIB	8%	Increased eccentricity observed	true positive

5.2. Performance Analysis

Quantification of anomaly detection performance in continuous time-series data can be viewed from at least two different perspectives. If the objective is to measure performance of a machine learning based method, one must tabulate *all predictions* (in our case every hour), we call it *sample-level performance* evaluation. However, in operational settings, plant operators are more interested in whether *all events* of interest (effects of which typically persist over longer periods) were successfully detected without resulting into many repeated alarms and false alarms, referred to as *event-level performance* evaluation. Here we present outcomes from both methods of performance quantification.

5.2.1. Coverage Period

In industrial applications data-driven models are typically trained on a subset of operational modes for an asset and hence only cover a subset of overall operational time period. If models are built to operate on very specific operational condition(s) out of many for an asset, the model output is not generated for untrained modes. For instance, a model trained on steady-state mode of a motor does not cover transient portions of data. For slow progressing failures, coverage on a subset of operational modes typically suffices, but in many cases, assets must be monitored or covered under all operational modes. Therefore, model coverage is yet another metric that must be tracked along with model accuracy. *Coverage* is defined as the percentage of time period for which model generates an output versus total period for which operational data were available from the plant. It must be further noted that time durations where the asset is known to be non-operational (e.g. during outages) model output should be ignored. Generally, in most implementations, models do not generate predictions when the asset is down, even though data acquisition systems may still be recording data. In-situ determination of periods where model coverage can be considered valid is important to properly identify to assess whether the model outputs should be accepted or ignored. Methods to determine covered periods is beyond the scope of this paper, but coverage is important to calculating performance metrics.

In our 10-month long data, there were 7322 hours of data available. Excluding periods of outages for this pump train, our model produced predictions for 6826 hours, i.e. a coverage of 93.23%. Total 496 hours were lost in outages (planned loss of 308 hrs (4.21%)) and unplanned loss of ~188 hrs (2.56%)). Performance metrics were calculated for the rest of the covered period.

5.2.2. Event-level Performance Evaluation

Event-level performance evaluation is summarized in Table 2, which was carried out to answer the specific questions that were important to the operational staff. For reference, Table 3 describes unplanned outage or power de-rate events (as labeled in Figure 3) that form the basis of ground truth determination of the plant conditions. Given sensor inputs used for the algorithm SME opinion was used to determine whether the model is applicable in detecting and determining the cause of unplanned outage event. In other words events that are not likely observable from chosen input sensor channels are excluded from the analysis and only outage U3 and other issues related to same system are considered meaningful for performance evaluation.

Table 2. Event-level performance evaluation summary.

Q1. Given data have *known* pump related events, have we discovered them and how many?
• Yes, all events were identified and verified with plant records (see Table 1). • True Detection: 13/13 • 4/13 were late detections from operational value perspective, i.e. not enough early warning period • False Detection: none
Q2. If discovered, can we calculate detection horizon (early warning) compared to when it was recorded in field?
• Yes, a minimum of 2 days and 18 hrs prior to when it was first detected in the field during routine rounding. Potentially, 5+ days of early warning (see next section).
Q3. Did we discover any events that were not documented *apriori*?
• No. While not known to us when data were originally provided, all alerts were successfully tallied subsequent to model building and evaluation with documented plant history.
Q4. How many alerts were *late positives* or *false positives,* if any?
• Oxygen Level Adjustment Events – technically speaking, detection is accurate (from an anomaly detection model's perspective), but alerts were not generated based on detecting low O2 levels, rather on detection of manual intervention (spike in O2 level). Since these won't be of any practical use in reducing costs and are considered late positives it was suggested these alerts be suppressed as these are non-pump related events.
Q5. Did we miss any events of interest (*false negatives*)?
• No. We define events of interest where the root cause of an unplanned outage was related to pump and adjacent subsystems to determine if the cause should have been covered by our model and if yes, did we detect it. • As shown in Table 3 below, only one outage was caused by a pump related issue and hence we don't expect our model to detect other unrelated outage root causes elsewhere in the plant.

Table 3. Analysis of known unplanned outages (*U_1-U_6 in* Figure 3) to assess missed detection.

Outage	Resolution	Applicable	Result
U1	Feedwater Heater related outage	NO	N/A
U2	Drain Cooler and FW Heater related	NO	N/A
U3	Feedwater Pump Vibration Issue and Repair	YES	Detected
U4	Main Steam Bypass Valve Malfunction	NO	N/A
U5	Heater Drain Valve Failure	NO	N/A
U6	Discharge Thermal Limits	NO	N/A

5.3. Sample-level Performance Evaluation

Sample by Sample accuracy evaluation computes machine learning algorithms' performance in terms of traditional metrics using confusion matrix. This requires labeling all time-series data points (samples) with a *ground-truth* label to compare with the predicted state. Such a ground-truth is hard to establish in real field data as there are several challenges described below:

- Failure reports document when a failure was detected in the field but determining the onset of preceding degradation that should have been detectable in data is a challenging task towards labeling the time series data. When precise knowledge is lacking, any assessment of the model's performance can, at best, provide indicative rather than precise results. In our specific scenario, we annotate the time series data using our understanding of known issues and the point at which their impact first became discernible in the data.
- In the initial phase of degradation, the symptoms are often intermittent. While it is important to detect as early as these symptoms first emerge, they may not always be consistently observable in periods following that detection, which affects this type of evaluation.
- There may be a number of confounding and overlapping periods where multiple issues/degradations may have been present. From detection point of view, a model should continuously flag the presence of degradation, however the root-cause may not appear consistent between those alerts. Detangling of these alerts is a manual process and can be hard to get it exact.

Keeping these challenges in mind we used the following approach for this evaluation. We labeled the time series with known anomalies based on consistency of the fault signature prior to an event. For example, if there is a bearing related fault observed through high vibrations, we use trends in vibration sensors to determine when they have statistically changed behaviors. Likewise, each known event can be characterized by corresponding best known signatures to label the data.

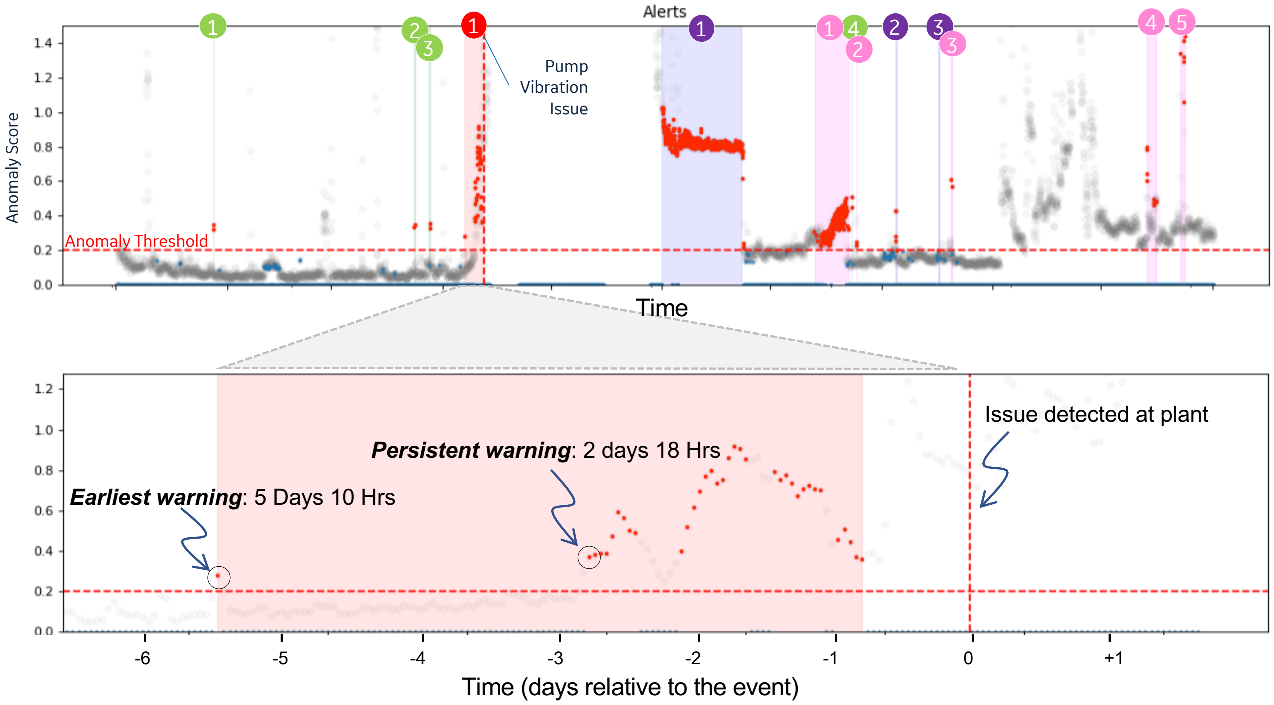

Figure 4. Trade-off between early warning vs. alert persistence condition: 5+ days for earliest indicator (no persistence) vs. 2+ days with a four-hour alert persistence.

Alert Persistence:

Noisy data and model errors can result into several predictions that cross anomaly threshold intermittently. Such prediction would result into false alerts unless high anomaly scores persist over longer periods. Persistence is indicative of systematic change in operational characteristics, often due to degradation or operational environment changes. Therefore, alert generation is a post-processing step, where based on downstream process and operator preference additional conditions must be satisfied before predictions are converted into an alert. Once such requirement is that of persistence, which requires a certain minimum number of predictions to persist above anomaly threshold before an alert is raised. It's often a trade-off between early warning and false positives.

Evaluation: As shown in Figure 4 the first alert with relevant signature (indicative of high vibrations) was briefly observed at about 5 days and 10 hours (t_1) prior to when it was detected at the plant. However, the alert subsided after that before becoming persistent about 2 days and 18 hours (t_2) prior to manual detection. For the purpose of evaluation, if we consider t_1 as the first desired indication of early warning, all subsequent data points must be labeled with *fault label* ground-truth. This results in a total of 995 hours that span anomalous time-series segments, if no persistence condition is applied. This would allow for an early warning for 5+ days for the key event of interest related to the FWP. However, with a persistence condition of 4 hours or more total anomalous segments reduce to 881 hours, and the best early warning horizon reduces to only 2+ days.

Table 4. Quantitative performance metrics.
N^{actual}, A^{actual} are number of nominal and anomalous labels, and
N^{pred}, A^{pred} are model predicted nominal and anomalous points.

Without a persistence logic: early warning of 5+ days

Hourly evaluation

	N^{pred}	A^{pred}	*totals*
N^{actual}	5823	8	*5831 hrs*
A^{actual}	245	750	*995 hrs*

Percentage hours

	N^{pred}	A^{pred}
N^{actual}	99.86%	0.14%
A^{actual}	24.62%	75.38%

With alert persistence of four hours: early warning of 2+ days

	N^{pred}	A^{pred}	*totals*
N^{actual}	5937	8	*5945 hrs*
A^{actual}	131	750	*881 hrs*

	N^{pred}	A^{pred}
N^{actual}	99.93%	0.07%
A^{actual}	14.86%	85.14%

6. CONCLUSIONS

As shown in evaluations above, the FWP early warning model is shown to provide warning with persistence at least 2 day and 18 hr. prior to when it was detected at the plant. Such warning could have prevented the catastrophic event and unplanned downtime through careful derating of the pump or switching to a redundant system. This supports O&M cost savings potential through detection and early warnings. It must be noted that more validation using other examples of similar issues may be valuable if data became available. The key learning experience from applying such an approach in real plant data was to tackle the challenges of obtaining a reasonable ground truth. While metrics like precision, recall, accuracy and confusion matrix can be used to optimize the algorithms, it was important to evaluate performance in terms of operational metrics relative to events of interest and whether those events were detected and with what lead times. Operational staff mainly cared about when a particular event was detected and if the algorithm provided explanations in terms of leading contributors to the anomaly scores. Some events (e.g. oxygen dosing) that were anomalous operations from algorithm's point of view were not interesting to the operational staff, which suggests that alert generation must be accompanied with post processing to suppress non-degradation related alerts, which is accomplished as a joint exercise with SMEs. Therefore, while sophisticated algorithms can be effective in industrial environments substantial domain knowledge is required to make these systems trustworthy and usable to realize meaningful cost savings.

6.1. Deployment

Work presented in this paper is still in proof of concept stages and far from deployment. However, the results obtained so far have been promising enough to the operational staff to allow continue investigation and further validation. Experience from other domains, where this modeling scheme has been successful deployed, and a positive outcome from this analysis thus far is encouraging and point to applicability of this approach for industrial environments in general. Future work plan has been put together towards maturing this method for a broader implementation in power plants.

6.2. Ongoing and Future Work

Work presented here continues in several directions to further validate our findings. For instance, on one side we are working on establishing reusability of model on a similar asset from another train in the plant. If the same model, as learned here, doesn't generalize as is we plan to investigate how to adapt/update to the new asset of similar kind through incremental training. On the other hand, we are also exploring if our algorithm can be used for different makes and models of the same asset as well as generalize to other types of assets at the plant through transfer learning. We are also working on expanding the model with chemistry parameters, to potentially detect deterioration in oxygen levels early on and use them as alerts to operators, rather than detecting the intervention as an anomaly. Finally, we plan to explore if expanding the model with parameters from the reactor core (e.g. reactor water level) can provide a more accurate assessment of operational mode of the plant.

7. ACKNOWLEDGEMENT

Authors would like to express gratitude to Constellation's plant staff, innovation teams for furnishing data and providing experiential domain knowledge. GE Hitachi's support with OEM design information, cost share and facilitating data sharing platforms. The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0001290. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

ACRONYMS

AI	Artificial Intelligence
BWR	Boiling Water Reactor
DOE	(US) Department of Energy
FWP	Feed Water Pump
INPO	Institute of Nuclear Power Operations
LCOE	Levelized Cost of Electricity
ML	Machine Learning
MSE	Mean Squared Error
NPP	Nuclear Power Plant
NRC	Nuclear Regulatory Council
OEM	Original Equipment Manufacturer
O&M	Operations and Maintenance
PHM	Prognostics and Health Management
SME	Subject Matter Expert
SMR	Small Modular Reactor (Nuclear)

BIBLIOGRAPHY

10 CFR 50.65 (NRC). (1974). *Requirements for monitoring the effectiveness of maintenance at nuclear power plants.* US Nuclear Regulatory Commission. US NRC.

Alex Tank, I. C. (2018). *Neural granger causality for nonlinear time series*. Retrieved from arXiv preprint: arXiv:1802.05842

Coble JB, G. L. (2013). Approaches to Quantify Uncertainty in Online Sensor Calibration Monitoring. *ANS Winter Meeting and Technology Expo.*

Coble, J., Ramuhalli, P., Bond, L. J., Hines, J., & Upadhyaya, B. (2015). A review of prognostics and health management applications in nuclear power plants. *Interational Journal of Prognostics and Health Management, 6*, 16.

Feng Xue, W. Y. (2020). Deep anomaly detection for industrial systems: a case study. *Annual Conference of the PHM Society.* PHM Society.

GE Hitachi. (2017). *BWRX-300*. Retrieved from https://nuclear.gepower.com/build-a-plant/products/nuclear-power-plants-overview/bwrx-300

Granger, C. W. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica, 37*(3), 424–438.

Hines, J. W., Coble, J., & Bailey, B. K. (2010). A Novel Method for Monitoring Single Variable Systems for Fault Detection, Diagnostics and Prognostics. *International Journal of Performability Engineering, 6*(5), 477.

Huang, H., & Kasiviswanathan, S. P. (2015). Streaming anomaly detection using randomized matrix sketching. *VLDB Endowment.*

Huang, H., Yan, W., Wang, T., & Xue, F. (2018). Imbalanced Time Series Classification with Nonlinear Causal Learning. *Woodstock '18:ACM Symposium on Neural Gaze Detection* (p. 9). New York: ACM.

Huang, H., Yoo, S., Yu, D., & Qin, H. (2015). Diverse Power Iteration Embeddings: Theory and Practice. *IEEE Transactions on Knowledge and Data Engineering*.

INPO. (2021). *Institute of Nuclear Power Operations*. Retrieved from Nuclear Costs in Context: INPO.info

Kaiming He, X. Z. (2016). Deep residual learning for image recognition. *IEEE conference on computer vision and pattern recognition* (pp. 770-778). IEEE.

Moleda, M., Momot, A., & Mrozek, D. (2020). Predictive maintenance of boiler feed water pumps using SCADA data. *Sensors, 20*(2), 571.

NEI. (2021). *Nuclear Costs in Context.* Retrieved from Nuclear Energy Insititute: https://www.nei.org/resources/reports-briefs/nuclear-costs-in-context

Ramuhalli, P., Walker, C., Agarwal, V., & Lybeck, N. (2021). Nuclear Power Prognostic Model Assessment for Component Health Monitoring. *12th ANS NPIC-HMIT*, (pp. 976-986).

Yadav, V., Agarwal, V., Gribok, A. V., Hays, R. D., Pluth, A. J., Ritter, C. S., . . . Iyengar, R. (2021). The State of Technology of Application of Digital Twins. *TLR/RES-DE-REB-2021-01* .

BIOGRAPHIES

Abhinav Saxena is a Principal Scientist in AI & Learning Systems at GE Research. Abhinav has been developing ML/AI-based PHM solutions for various industrial systems (aviation, nuclear, power, and healthcare) at GE and has been driving integration of AI-based PHM analytics in GE's industrial systems. He is the PI for ARPA-E GEMINA program led by GE Research on AI-Enabled Predictive Maintenance Digital twins for Advanced Nuclear Reactors. Abhinav is also an adjunct professor in the Division of Operation and Maintenance Engineering at Lulea˚ University of Technology, Sweden. Prior to GE, Abhinav was a Research Scientist with SGT Inc. at NASA Ames Research Center for over seven years. Abhinav's interests lie in developing PHM methods and algorithms with special emphasis on deep learning and data- driven methods in general for practical prognostics. Abhinav has published over 100 peer reviewed technical papers and has co-authored a seminal book on prognostics. He actively participates in several SAE standards committees, IEEE prognostics standards committee, and various PHM Society educational activities, and is a Fellow of the PHM Society. He also served as chief editor of International Journal of Prognostics and Health Management between 2011-2020. Abhinav actively participates in organization of PHM Society conferences and various AI workshops on topics of Digital Twins and AI in Industrial applications.

Helena Goldfarb is a Senior Scientist in AI & Learning Systems at GE Research. Helena has been developing ML/AI-based solutions for various industrial systems with GE businesses and Lockheed Martin. She also has been working on programs sponsored by the government agencies, such as Joint Strike Fighter Propulsion program and DARPA's Measuring Biological Aptitude program. Helena's interests lie in developing ML-based industrial PHM solutions to drive carbon free energy generation.

Jeffrey S. Clark, PE is a Senior Staff Engineer in the Corporate Equipment Reliability Group at Constellation. Constellation is the largest US operator of nuclear plants accounting for 25% of nuclear power in the US. Jeff's 34 years in the nuclear field include design, plant systems, programs, training and project management. In his current role Jeff is the lead for a Constellation fleet project to identify single point vulnerabilities where equipment failure could lead to a loss of power generation. The project also includes assessment of equipment failure risk and reviewing potential maintenance, operational and monitoring strategies for implementation to mitigate generation risk. Risk and resource optimization have always been his passions. Jeff has a BS in Civil/Structural Engineering from the University of Illinois in Champaign-Urbana, IL and is a registered Professional Engineer in IL & GA. Jeff was previously the Exelon fleet lead for the Fukushima Seismic project also leading weekly industry calls and a member of the TIP award winning SSOT group with NEI & EPRI