INL/RPT-22-68942
Revision 0

Light Water Reactor Sustainability Program

Technical Basis for Advanced Artificial Intelligence and Machine Learning Adoption in Nuclear Power Plants

September 2022

U.S. Department of Energy

Office of Nuclear Energy

INL/RPT-22-68942
Revision 0

Technical Basis for Advanced Artificial Intelligence and Machine Learning Adoption in Nuclear Power Plants

Vivek Agarwal, Cody Walker, Koushik Manjunatha, Torrey Mortenson, Nancy Lybeck, Anna Hall, Rachael Hill, and Andrei Gribok

September 2022

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy

EXECUTIVE SUMMARY

The research and development reported here is part of the Technology Enabled Risk-Informed Maintenance Strategy project sponsored by the U.S. Department of Energy's Light Water Reactor Sustainability program. The primary objective of the research presented in this report is to produce a technical basis for developing explainable and trustable artificial intelligence (AI) and machine learning (ML) technologies. The technical basis will lay the foundation for addressing the technical and regulatory adoption challenges of AI/ML technologies across plant assets and the nuclear industry at scale and to achieve seamless cost-effective automation without compromising plant safety and reliability.

The technical basis ensuring wider adoption of AI/ML technologies presented in this report was developed by Idaho National Laboratory (INL), in collaboration with Public Service Enterprise Group (PSEG) Nuclear, LLC. To develop the initial technical basis, the circulating water system (CWS) at the PSEG-owned plant sites was selected as the identified plant asset. Specifically, the issue of waterbox fouling diagnosis in the CWS using different types of CWS data is presented to address the said challenge. The approach presented in this report is based on the closed–loop forward–backward process that tries to capture the advancements in data science addressing the explainability of AI/ML outcomes, user-centric interpretability of those outcomes, and how user interpretation can be used as feedback to further simplify the process. A prototype interface is developed to present a focused component-level display of the ML model outputs in a usable and digestible form.

The forward process moves from data-to-decision and is one of the main parts of research performed as AI/ML technologies are developed. The forward process entails a rigorous mathematical approach that accounts for data preprocessing; data integration; transformation of the data into usable information to train, validate, and test ML models; hyperparameter optimization of ML models; uncertainty quantification of final outputs by accounting for the accumulation of errors; and presentation of the results to the end-user. The focus is to explain AI/ML solutions by utilizing objective metrics, such as Local Interpretable Model-agnostic Explanations and Shapley Additive Explanations, that can capture the rigorous mathematics. The focus of a metric-based approach is to quantify the effectiveness of the explanation based on performance differences between the ML models, the number of features used to construct the explanation, and the stability of the explanation.

In the backward process, the objective metrics developed as part of the forward process to explain AI/ML technologies are verified by the end-user. As part of the backward process, a user-centric visualization is developed to present AI/ML outcomes with objective metrics and other information to elicit user interpretation. Based on elicited input from end-users with different levels of expertise and functional positions within the organization, the objective metrics and visualization are adapted to ease the user interpretation and requirements of AI/ML outcomes to inform decision.

The technical basis developed in this report will be extended to other fault modes and to a wide user verifiability study. For the next year, research will focus on the verification and validation of explainability and trustworthiness of AI/ML technologies. As part of the path forward, novelty detection—where the current data are compared to the training data set to identify operating regimes outside the scope of the model—will be explored in detail to determine its role in improving ML trustability.

ACKNOWLEDGEMENTS

This report was made possible through funding from the U.S. Department of Energy (DOE)'s Light Water Reactor Sustainability program. We are grateful to William Walsh of DOE and Bruce P. Hallbert and Craig A. Primer at Idaho National Laboratory (INL) for championing this effort. We thank Kelsey B. Gaston at INL for the technical editing of this report. We thank Barry Pike III and Lauren M. Perttula of RED, Inc. for some of the graphics contained in the report. We would also like to thank Matthew Pennington and Lee Papasergi at the PSEG Monitoring and Diagnostic Center for their valuable technical contributions.

CONTENTS

EXECUTIVE SUMMARY iii

ACKNOWLEDGEMENTS iv

ACRONYMS x

1. INTRODUCTION AND MOTIVATION 1

2. APPROACH ENSURING ADOPTION OF ARTIFICIAL INTELLIGENCE BASED SOLUTIONS IN NUCLEAR INDUSTRY 2

3. DATA-TO-DECISION CONSIDERATION ENSURING ADOPTION OF ARTIFICIAL INTELLIGENCE BASED SOLUTIONS 4
   - 3.1 Waterbox Fouling Issue in Circulating Water System 5
   - 3.2 Data and Variabilities 5
   - 3.3 Feature Explainability Metrics 6
     - 3.3.1 Shapley Additive Explanations (SHAP) 6
     - 3.3.2 Local Interpretable Model-agnostic Explanations (LIME) 8
   - 3.4 Machine Learning Models 8
     - 3.4.1 Extreme Gradient Boosting 8
     - 3.4.2 Random Forest 8
     - 3.4.3 Deep Neural Network 9
     - 3.4.4 Hyperparameters of ML Models 9
   - 3.5 Results 10
     - 3.5.1 XGBoost Performance 10
     - 3.5.2 Random Forest and Deep Neural Network Performance 16

4. HUMAN FACTORS CONSIDERATIONS TO EVALUATE A USABLE ADOPTION OF ARTIFICIAL INTELLIGENCE BASED SOLUTIONS 20
   - 4.1 Introduction 20
     - 4.1.1 Key Concepts: Explainability and Trust 20
     - 4.1.2 Trust in Automation 20
     - 4.1.3 Explainable AI 22
     - 4.1.4 Human-Centered Artificial Intelligence 23
     - 4.1.5 Nuclear Safety Culture 25
     - 4.1.6 Assessing Trust and Explainability in Interface 25
   - 4.2 Method 26
     - 4.2.1 Experimental Design 26
     - 4.2.2 Interface Design 26
     - 4.2.3 Procedure 28
   - 4.3 Results 28
     - 4.3.1 Results of Maintenance Decision Task 28
     - 4.3.2 Interface Design and Model Feedback 29
   - 4.4 Discussion 30
     - 4.4.1 Potential Barriers to AI/ML Adoption in the Nuclear Industry 31
     - 4.4.2 Increasing Trust in AI 31
     - 4.4.3 Matching User Mental Models 32

4.5 Potential Research Gaps 32
4.5.1 Explainability and Trust Construction 32
4.5.2 Deeper Interface Development to Support Diagnostic Processes 32
4.5.3 Matching Model Process to User's Mental Models 33
4.5.4 Nuclear Safety Culture Challenges 33

5. SUMMARY AND PATH FORWARD 33

6. REFERENCES 34

FIGURES

Figure 1. Transition from a PM program to a risk-informed PdM program. 1

Figure 2. Data-to-decision roadmap for a risk-informed PdM strategy. 3

Figure 3. Forward-backward process to ensure reproducibility and interpretability of AI/ML technologies. 3

Figure 4. An example of changes to the CWS process data before and after waterbox fouling at the Salem's Unit 2 waterbox 22B 5

Figure 5. Examples of CWS measurements: showing Gross load, temperatures for the Stator, MIB, MOB, Pump status, DT, and Motor current. 7

Figure 6. Feature value influence on SHAP values in prediction of Waterbox fouling. 11

Figure 7. Feature importance for all the features in predicting waterbox fouling versus healthy data. 11

Figure 8. Local interpretation of an instance corresponding to waterbox fouling. 12

Figure 9. Local interpretation of an instance corresponding to healthy condition. 12

Figure 10. Feature value influence (when feature DT is missing) on SHAP values in prediction of Waterbox fouling. 13

Figure 11. Feature value influence (when feature Motor Current is missing) on SHAP values in prediction of Waterbox fouling. 13

Figure 12. Feature value influence (when feature MOB temperature is missing) on SHAP values in prediction of waterbox fouling. 14

Figure 13. A test instance (2018-02-02 20:00:00) corresponding to waterbox fouling when all the features are available. 15

Figure 14. A test instance (2018-02-02 20:00:00) corresponding to waterbox fouling when feature DT is missing. 15

Figure 15. A test instance (2018-02-02 20:00:00) when feature Motor Current is missing. 16

Figure 16. Training and test data were split into time segments with similar amounts of labeled data for healthy and waterbox fouling data. These time segments became folds for cross validation of the models' performance. 17

Figure 17. MOB temperature is showing a seasonal dependance as the temperatures are higher during the summer and lower during the winter. 18

Figure 18. Local explanation of the RF's waterbox fouling prediction using LIME. Positive, green values are those contributing to waterbox fouling, while negative, red values contribute to a healthy determination. The RF has 82% confidence in this prediction. Features have been anonymized. 19

Figure 19. Relationship between calibration, resolution, and automation capability. Reproduced from Lee and See [18]. 22

Figure 20. Examples of tasks that fall under two-dimensional HCAI. 24

Figure 21. Scenario selection screen. 26

Figure 22. Interface with location markers. 28

Figure 23. Graphical representation of a model being fit within the training distribution (white area) and how the extrapolation into new domains (gray area) may vary. Figure from [48]. 34

TABLES

Table 1. DNN parameters. 9

Table 2. Classification performance results. (All the features are available). 10

Table 3. Prediction performance and feature importance under unavailability of different measurements during training. 14

Table 4. SHAP model prediction results for the CWP 13B instance 2018-02-02 20:00:00 for Waterbox fouling. 16

Table 5. RF training and testing accuracies with inputs of DT, MOB temperature, and motor current. The highlighted portion shows an instance of testing data well within the training distribution and an instance of overfitting. 17

Table 6. Prediction performance with unavailability of different features during training. 18

Table 7. Global feature importance for DNN and RF. 18

Table 8. Study design. 26

Table 9. Interface features and locations. 27

Table 10. Results of decision task. 29

ACRONYMS

AI	artificial intelligence
CWP	circulating water pump
CWS	circulating water system
DOE	Department of Energy
DNN	deep neural network
DT	differential temperature
HCAI	human-centered artificial intelligence
INL	Idaho National Laboratory
LIME	Local Interpretable Model-agnostic Explanations
LWR	light water reactor
LWRS	Light Water Reactor Sustainability
M&D	monitoring and diagnostic
MIB	motor inboard
MOB	motor outboard
ML	machine learning
NN	neural networks
NPP	nuclear power plant
NRC	Nuclear Regulatory Commission
O&M	operation and maintenance
PdM	predictive maintenance
PM	preventive maintenance
PSEG	Public Service Enterprise Group
RF	random forest
R&D	research and development
SHAP	Shapley Additive Explanations
TERMS	Technology-Enable Risk-informed Maintenance Strategy
XAI	explainable artificial intelligence

TECHNICAL BASIS FOR ADVANCED ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING ADOPTION IN NUCLEAR POWER PLANTS

1. INTRODUCTION AND MOTIVATION

The primary objective of the research presented in this report is to produce a technical basis for developing explainable and trustable artificial intelligence (AI) and machine learning (ML) technologies. The technical basis will lay the foundation for addressing the adoption challenges of AI/ML technologies across plant assets and the nuclear fleet to achieve risk-informed predictive maintenance (PdM) strategies at commercial nuclear power plants (NPPs). Over the years, the nuclear fleet has relied on labor-intensive, time-consuming preventive maintenance (PM) programs, driving up operation and maintenance (O&M) costs to achieve high-capacity factors. A well-constructed, risk-informed PdM approach for an identified plant asset has been developed in [1] that takes advantage of advancements in data analytics, AI/ML, physics-informed modeling, and visualization. These technologies would allow commercial NPPs to reliably transition from the current labor-intensive PM programs to a technology-driven PdM program (see Figure 1), thus eliminating unnecessary O&M costs.

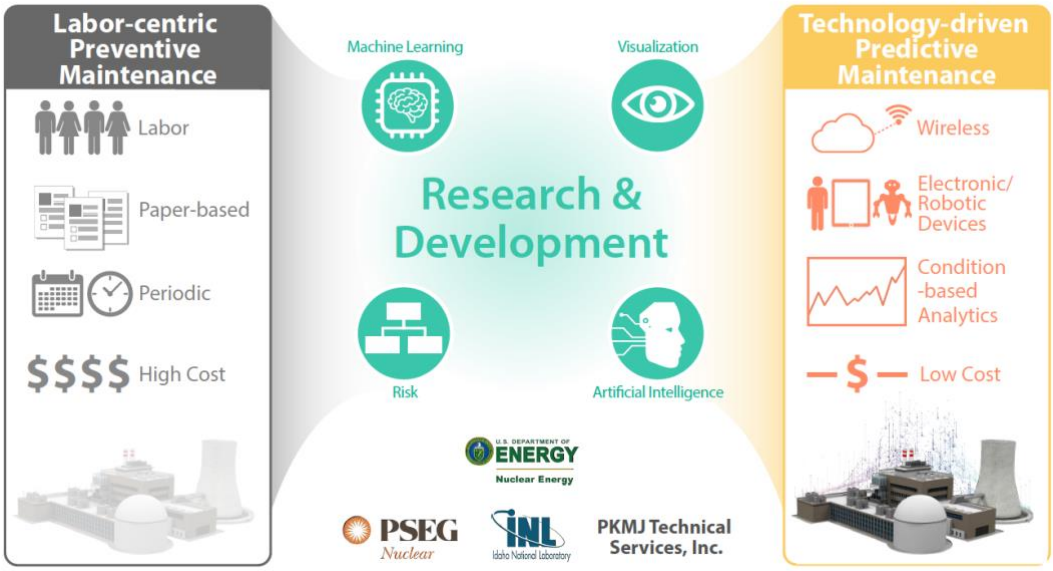

Figure 1. Transition from a PM program to a risk-informed PdM program.

The research and development (R&D) reported here is part of the Technology Enabled Risk-Informed Maintenance Strategy (TERMS) project sponsored by the U.S. Department of Energy (DOE)'s Light Water Reactor Sustainability (LWRS) program. The LWRS program is an R&D program conducted in close partnership with industry to provide the technical foundations for licensing, managing, and economically operating the current fleet of NPPs. To achieve both program and pathway goals [4], a series of pilot projects are underway to develop and demonstrate new technologies that can affect transformative change in NPP operations and support. The TERMS pilot project on risk-informed PdM strategy is developing the necessary technologies and methodology to address the five challenges outlined below

There are several challenges with research, development, demonstration, and deployment that need to be addressed as plants transition from a PM program to a risk-informed PdM program. These include:

- Integration and synchronization of data collected at different spatial and temporal resolutions over several decades
- Concerns over data imbalance (i.e., a significant amount of data collected when the plant system is operating under normal condition with no degradation and a limited amount of data are associated with degradation of the plant system)
- Scalability of the AI/ML technologies across plant systems and the nuclear fleet to meet current and future application-specific requirements
- User-centric visualization scheme to enable cross-facility users to understand the states of plant systems and the plant itself without having to remember and use separate visualization software
- Explainability and trustworthiness of AI/ML technologies enabling modernization and automation across the plant.

The research addressing the first four challenges is covered in [1–3]. The challenge of explainability and trustworthiness of AI/ML technologies is critical for addressing their technical and regulatory adoption across the nuclear industry at scale and achieving seamless cost-effective automation without compromising plant safety and reliability. The technical basis ensuring wider adoption of AI/ML technologies presented in this report was developed by Idaho National Laboratory (INL), in collaboration with Public Service Enterprise Group (PSEG) Nuclear, LLC. To develop the initial technical basis, the circulating water system (CWS) at the PSEG-owned plant sites were selected as the identified plant asset. Specifically, the issue of waterbox fouling diagnosis in the CWS using different types of CWS data is presented to address said challenge. The approach presented in this report is based on the closed–loop forward–backward process that tries to capture the advancements in data science addressing the explainability of AI/ML outcomes, user-centric interpretability of those outcomes, and how user interpretation can be used as feedback to further simplify the process. A prototype interface is developed to present a focused component-level display of the ML model outputs in a usable and digestible form.

This report is organized as follows: Section 2 presents the forward–backward closed-loop approach to enhance the possibility of AI/ML adoption in the nuclear industry and describes the fundamental aspects of the initial technical basis discussed in this report. Section 3 uses scenarios to demonstrate how to improve the interpretability of AI/ML technologies using metrics as the time-series data evolves over time, specifically, when data are missing or when new data are added to the already trained ML model. Section 4 describes the creation and evaluation of the interface with key considerations of explainability and trust. The testing of the developed interface presenting the AI/ML outcomes in a semi-structured interview format was performed with monitoring and diagnostic (M&D) analysts serving as participants. Section 5 summarizes research progress and discusses the path forward by highlighting a few open technical challenges.

2. APPROACH ENSURING ADOPTION OF ARTIFICIAL INTELLIGENCE BASED SOLUTIONS IN NUCLEAR INDUSTRY

Transforming the embedded knowledge in heterogeneous data sources collected by NPPs across structures, systems, and components into usable information for decision-making by the human-in-the-loop requires a systematic approach. The approach might include the use of AI/ML technologies, but always follows these basics: (1) identification of plant asset of interest; (2) collection of all the relevant data associated with the identified plant system; (3) pre-processing of the data; (4) extraction of salient information from the data; (5) inputting the information and data into a model (which could be a physics-based model, stochastic model, deterministic model, or AI/ML model); (6) model outputs; (7) visualization of outputs to be presented to the user; and (8) finally, decision or action taken by the user.

An example of a data-to-decision roadmap developed for the risk-informed PdM strategy is shown in Figure 2. Observe, in Figure 2, the input layer presents different data types that is related to the CWS of a

PSEG-owned power plant and be true for any plant system. The inputs are broadly categorized into constant data values, periodic data, and real-time data. The details of real-time data are presented in Section 3. The risk-informed predictive analytics provide insight into how these different data types are used for different purposes and integrated. Different forms of analysis performed by models and their integration could lead to different outputs. The integration of these model outputs needs to be simple, clear, and consistent, ensuring the resultant outputs are reproducible and interpretable by the end-user.

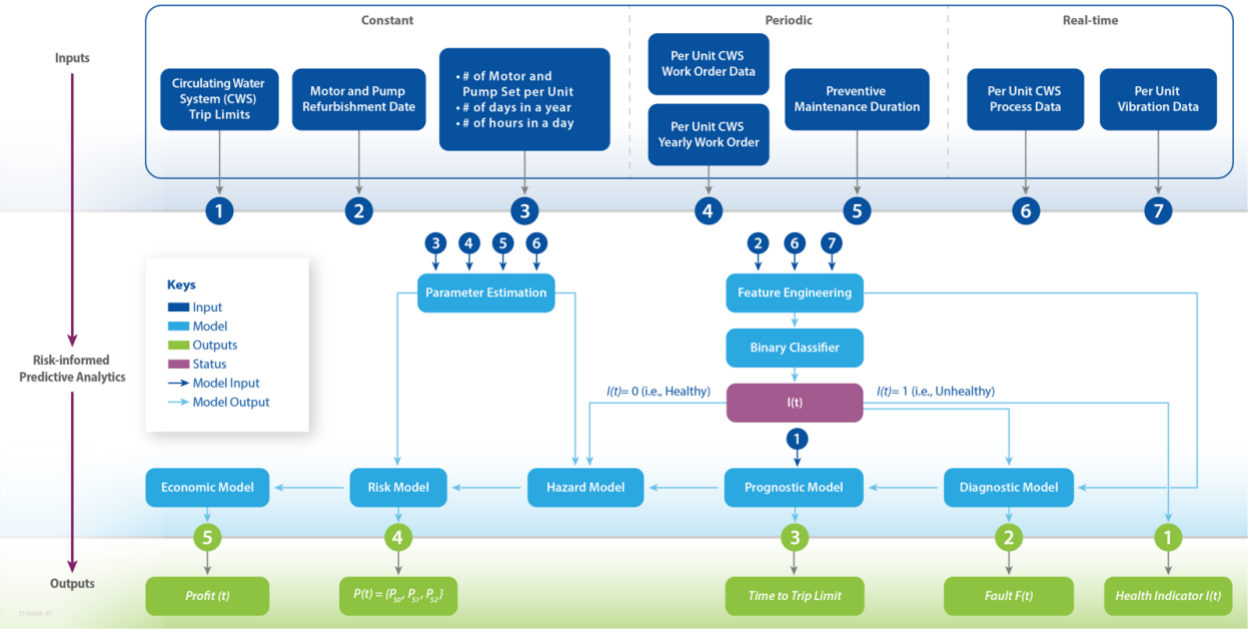

Figure 2. Data-to-decision roadmap for a risk-informed PdM strategy.

To ensure the data-to-decision roadmap, shown in Figure 2, is widely adopted by the nuclear industry to address technical and regulatory concerns, automation with a human-in-the-loop, forward–backward closed–loop process can be implemented to ensure consistent results that are interpretable for the end-users, as shown in Figure 3.

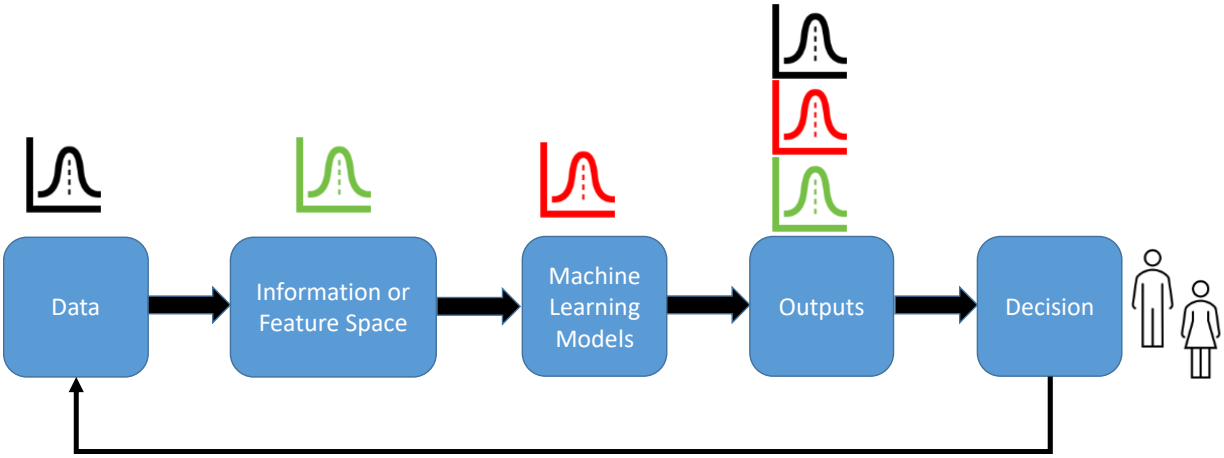

Figure 3. Forward-backward process to ensure reproducibility and interpretability of AI/ML technologies.

The forward process moves from data-to-decision (as in Figure 2) and is primary focus of research performed by data scientists as they develop AI/ML models. The forward process entails a rigorous

mathematical approach that accounts for: data preprocessing; data integration; transformation of the data into usable information to train, validate, and test ML models; hyperparameter optimization of ML models; uncertainty quantification of final outputs by accounting for the accumulation of errors; and presentation of results to the end-user (as shown in Figure 3). In the forward process, without considering the end-user, there are variabilities introduced due to data, information, and models, that influence outputs. These variabilities include:

- Data unavailability, inclusion of new data streams, changes in operating conditions, season variations, and others
- Feature importance for a particular fault mode
- Generalizability of linear and non-linear ML models (for example, neural networks, support vector machines, kernel regressions, random forest (RF), convoluted neural network, etc.), optimization of hyperparameters, and choice between supervised, semi-supervised, and unsupervised ML approaches.

One of the aspects that also needs to be considered in the forward process is the need to retrain, revalidate, and retest ML models when the above-mentioned variabilities occur in the current data the extent (as determined by the novelty detection) that the base ML model performance needs to be re-evaluated. It is important to evaluate these variabilities and determine a strategy to re-evaluate the base ML model. The re-evaluation strategy could be either periodic, on-demand, performance-based, or trigger-based based considering the on percentage change in data.

The focus of this report is to explain AI/ML solutions by utilizing objective metrics that can capture the rigorous mathematics. The use of a metric-based approach allows quantification of the effectiveness of the explanation based on performance differences between the ML models, the number of features used to construct the explanation, and the stability of the explanation.

The output values generated at the end of the forward process are expected to be used by an end-user in decision-making even if they are not readily and fully understood. This disconnect is expected to hurt the adoption of AI/ML technologies in critical decision-making. To address the technical communication barriers between data scientists and end-users, the backward process has been developed as a bridging approach. In the backward process section of this report, the objective metrics developed as part of the forward process to explain AI/ML technologies are verified by the end-user. Note, that though the end-user verifies the objective metrics, care is taken to ensure that the verification doesn't suffer from confirmation bias, as pointed out in [5]. A part of the backward process, a user-centric visualization is developed to present AI/ML outcomes with objective metrics and other information, to elicit user interpretation. Input elicited from end-users with different levels of expertise and functional position within the organization will be used to adapt the objective metrics and visualization to ease the user interpretation and requirements of AI/ML outcomes to inform the decision.

In the following sections, the initial approach taken to build a foundation of forward-backward closed-loop process for waterbox fouling in a CWS is presented and discussed.

3. DATA-TO-DECISION CONSIDERATION ENSURING ADOPTION OF ARTIFICIAL INTELLIGENCE BASED SOLUTIONS

This section will discuss the forward process of how data is used to train, validate, and test three ML algorithms for the case study problem of waterbox fouling. This section details the recorded data, its variabilities, and the steps used to condition the data. Cleaned data is then used in three ML algorithms (extreme gradient boosting, RF, and deep neural networks) for classification of waterbox fouling. To improve the explainability of these models, Shapley Additive Explanations (SHAP) [6] and Local Interpretable Model-agnostic Explanations (LIME) [7] were implemented. These methods detail each feature's contribution to the ML outcome. Improved explainability of how the model reached its conclusion is critical for acceptance and implementation of the algorithms.

3.1 Waterbox Fouling Issue in Circulating Water System

Waterbox fouling is a common maintenance issue at the PSEG-owned Salem NPPs. Fouling of the waterboxes typically occurs due to the accumulation of grass/debris in the waterbox, thus resulting in condenser tube blockage and reduced circulator water flow. This is a unique and frequent issue where the circulating water pump (CWP) intake comes directly from the river, resulting in a significant quantity of grass/debris. Primary symptoms of waterbox fouling include:

- Motor current increase (Sometimes seen by motor current decrease, but not often)
- Inlet pressure increase
- Waterbox differential temperature (DT) increase
- Condenser thermal performance loss.

Figure 4 shows an instance of waterbox fouling diagnosed in the Salem Unit 2, CWP 22B. An upward drift in DT and motor current was identified on July 23, 2018. Consequently, the gross load started dipping. Note in Figure 4, the CWP 22B motor current increased from 231 to 245 amps, and DT increased from 14°F to 16°F with Gross load not trending as expected. Motor current and DT decreased to 220 amps and 14°F, respectively, following waterbox cleaning on August 25, 2018, resulting in a 30-40 MWe improvement in gross load. The waterbox fault and approximate date of the shutdown were found by searching the Work Order and narrative log information provided by PSEG [1].

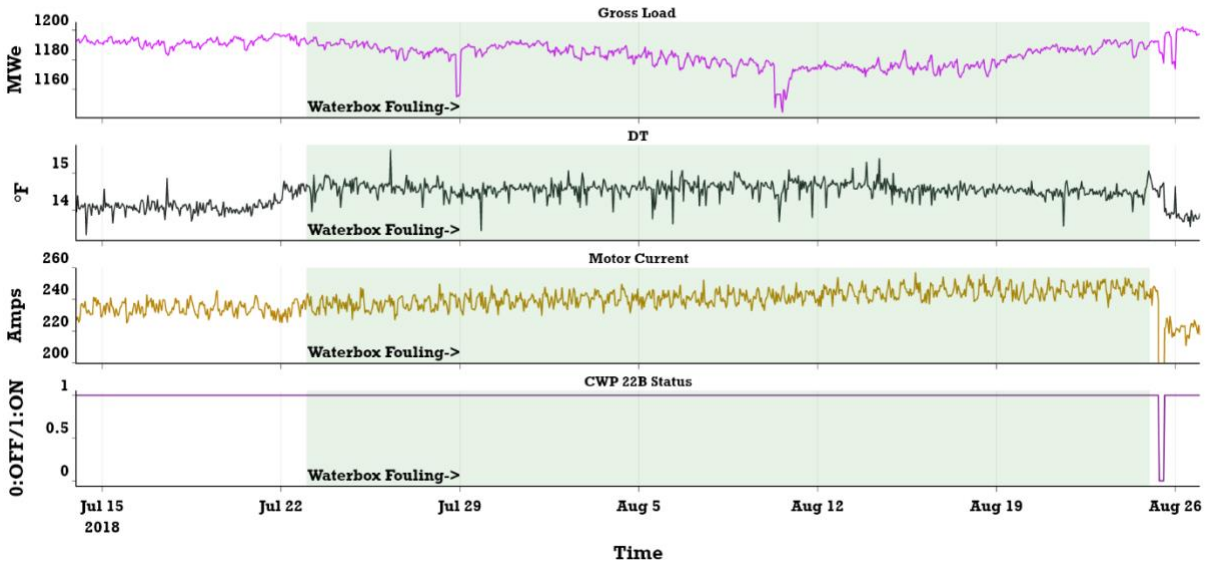

Figure 4. An example of changes to the CWS process data before and after waterbox fouling at the Salem's Unit 2 waterbox 22B.

3.2 Data and Variabilities

The Unit 1 and Unit 2 CWS process data are collected once every minute and stored in the Salem plant's OSI PI system. Due to file size restrictions, the project team received hourly CWS process data for both units, from 2009 to January 2021. The process data includes:

- Gross load (MWe)
- River level (ft)

- Ambient air temperature (°F)
- CWP inlet river temperature (°F)
- CWP outlet water temperature (°F)
- CWP motor status (ON or OFF)
- CWP motor stator winding temperature (°F)
- CWP motor inboard-bearing (MIB) temperature (°F)
- CWP motor outboard-bearing (MOB) temperature (°F)
- CWP motor current (Amps).

Detailed information about all the data types can be found in our previous work [1-3]. Figure 5 shows a sampling of CWS process data for a Salem unit.

Even though the data is available from 2008, not all the measurements were collected since that time. The motor current measurement is available from October 2017 onwards, and the online vibration measurements were collected from 2020 onwards. Other parameters, such as inlet pressure, were measured semi-periodically by hand. Due to the infrequent nature of those measurements, some of these parameters were excluded from the analysis. Additionally, the data collected had outliers, missing values, bad inputs, duplicate timestamps, daylight saving entries, and text entries. The outliers were identified using standard deviation and median filtering approaches. The repeated timestamps and daylight savings entries were removed and replaced with the missing timestamps. For missing values, a moving average approach was considered with a window size of 500 to 700 hours to fix continuous missing values. Some parameters, such as CWP MOB temperature, showed a seasonal dependance. In this situation, the parameter may contain more information about the ambient temperature than the condition of the motor. More details about the data cleaning can be found in our previous works [1], [3].

From the CWS associated plant process data, the following features are extracted for each pump-motor set:

- DT, calculated as the difference between the inlet river temperature and the outlet waterbox temperature. The outlet temperature is the combination of multiple pumps because each waterbox is connected to multiple CWPs.
- Inboard, outboard, and stator motor temperatures.
- Motor currents (available after September 2017).

3.3 Feature Explainability Metrics

This section covers the SHAP and LIME feature explainability metrics. These metrics are used to provide an explanation of how each feature contributes to the black-box algorithms' outputs. For a single data snapshot, these methods can be used to explain an individual feature's contribution. When used for an entire data set, these methods can be used to describe the feature's importance. This section details how SHAP and LIME are calculated.

3.3.1 Shapley Additive Explanations (SHAP)

SHAP values are a feature additive approach where the output is a linear combination of inputs [6]. SHAP quantifies the influence that each input feature has on the prediction output. Calculating the exact Shapley values is generally infeasible but calculating the approximate SHAP values can be accomplished with an explanatory model. For an input sample of x with M features, the prediction for j^{th} class is a linear combination of all the associated SHAP values:

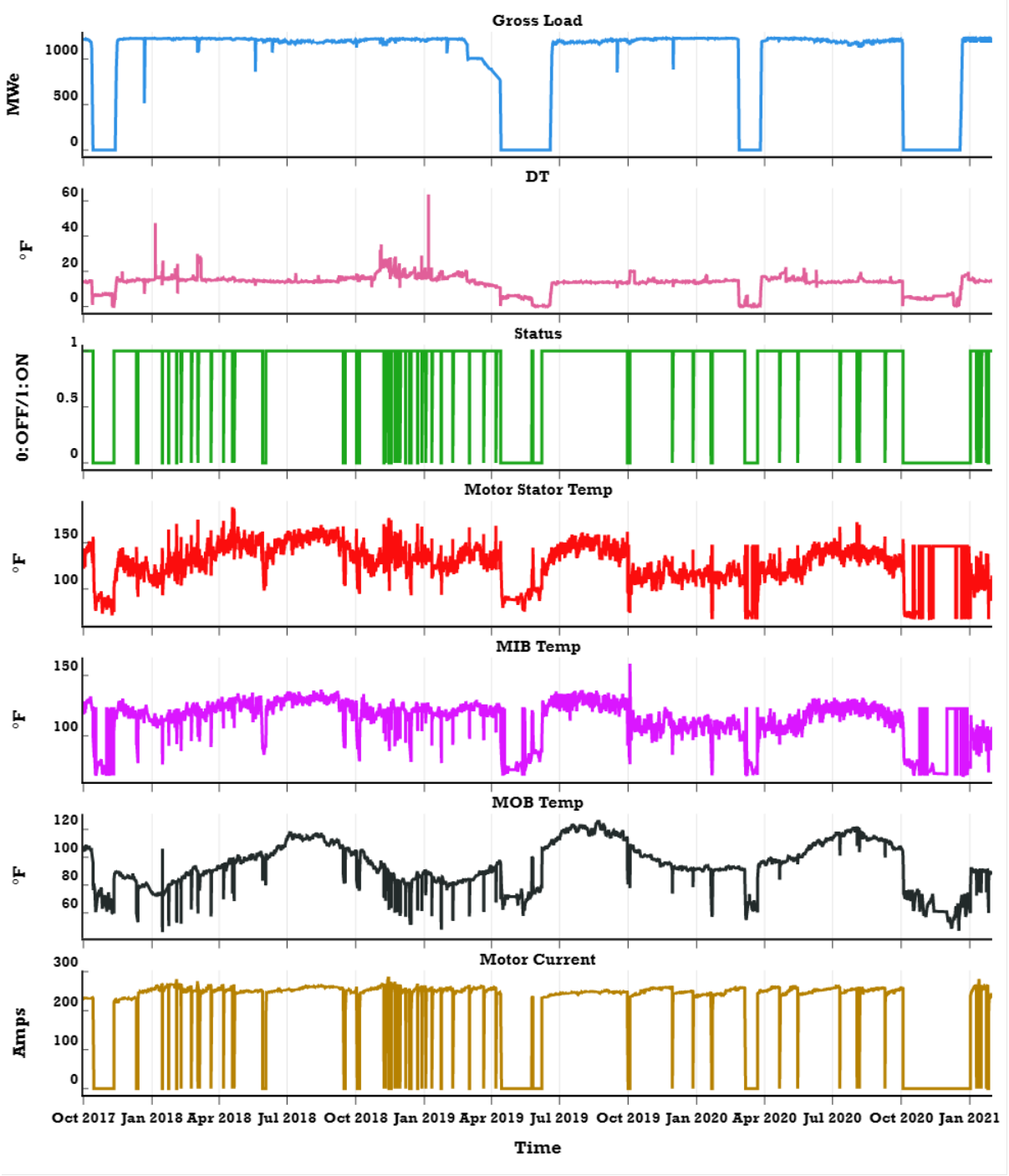

Figure 5. Examples of CWS measurements: showing Gross load, temperatures for the Stator, MIB, MOB, Pump status, DT, and Motor current.

$$f(x)_j = E\left[f(x)_j\right] + \sum_{i=1}^{M} \phi_{ij} \quad (1)$$

where $f(x)_j$, $E[f(x)_j]$, and ϕ_{ij} are the output (a.k.a. the logit) value for a fault, the average prediction value for a fault, and associated SHAP value for each feature, respectively. According to Equation (1), a

positive SHAP value associated with a feature indicates the feature is contributing to the prediction of j^{th} class label, whereas a negative SHAP value, indicates the feature is contributing to the prediction of a different class, $k \neq j$.

3.3.2 Local Interpretable Model-agnostic Explanations (LIME)

LIME attempts to identify an interpretable model that is locally faithful to the classification model in question [7]. The interpretable explanations need to be presentable and digestible to the human-in-the-loop. The interpretable model is a relatively straightforward model, such as a linear model or decision tree that tries to explain the feature contribution for a single point in time. The explanation, g, from this model determines the absence or presence of the interpretable components. These interpretable components are problem-dependent and will be different for numerical data, text, and images. Not all features, transformations, or combinations of features are readily explainable, so LIME attempts to ensure that the output is reasonably simple by introducing a measure of complexity, denoted as Ω(g). This measure of complexity is determined by the class of interpretable models, denoted as G. For example, Ω(g) can be the depth of the decision tree or the number of non-zero weights for the linear model. $L(f, g, \pi_x)$ is the measure of how well that model explanation, g, is approximating the probability of x belonging to a certain class, $f(x)$, in the locality defined by π_x. The LIME is produced by the following,

$$\xi(x) = \ argmin_{g \epsilon G}\ L(f, g, \pi_x) + \Omega(\mathrm{g}) \tag{2}$$

where the locality-aware loss function $L(f, g, \pi_x)$ is minimized with consideration of the explanation's complexity, Ω(g). This ensures that the result is locally faithful to the classifier and interpretable. Since no assumptions are made about the f, LIME outputs are model-agnostic. In Equation (2), the locality, π_x, is weight with a distance function such as an exponential kernel as follows,

$$\pi_x(z) \ = \ exp(-D(x, z)^2/\sigma^2) \tag{3}$$

where D is the distance function, x is the point of interest, z is randomly sampled from the feature space, and σ is the width. Although LIME is quite useful, it is limited by the class of interpretable models, G, and may not be useful if the underlying model is highly non-linear even in local predictions.

3.4 Machine Learning Models

The extracted parameters from plant process data form a feature vector $x \in X$, where X is a feature set of size $n \times m$. Every feature vector is associated with a class label $y \in Y$, where Y is label vector of length n. The parameters n and m are the number of feature vectors and the number of features in a feature vector, respectively. Three of the most common advanced ML models—eXtreme Gradient Boosting (XGBoost) [8], RF [9], and Deep Neural Network (DNN) [10]—were considered for predicting waterbox fouling. The details on each of the models are discussed in the following subsections. Further, the hyperparameters, which control the prediction performance of each model are briefly discussed.

3.4.1 Extreme Gradient Boosting

Gradient boosting refers to a class of ensemble ML algorithms that can be used for classification and regression. A special class of ensemble learning is boosting in which decision trees are added one at a time to the ensemble and fit to correct the prediction errors made by prior models. Such boosting models fit using a differentiable loss function and gradient descent optimization are called gradient boosting. XGBoost [8] is an open-source implementation of a gradient boosting algorithm.

3.4.2 Random Forest

RF is also a ML algorithm that can be used for classification and regression through the ensemble of decision trees [9]. While a single decision tree can be subject to variability and noise, RF overcomes these issues by generating many decision trees and bootstrapping samples from the training data. Bootstrapping involves random sampling with replacement, so each individual decision sees a slightly different training

data set. The RF then aggregates the outputs from the many decision trees to produce the final regression value or classification. RF excels because it is easy-to-implement, accurate, and has built-in feature importance metrics due to the structure of the decision trees. RF is readily available in python's sklearn library.

3.4.3 Deep Neural Network

A DNN contains multiple layers of interconnected neurons that create a complex, non-linear mapping from input to output [10]. The mapping is fit to the data through a process called backpropagation, where the neuron's weights are adapted to best fit the training data. DNN has been used successfully for a wide variety of applications including classification, natural language processing, and speech recognition. TensorFlow was the open-source, deep-learning library used in this research.

3.4.4 Hyperparameters of ML Models

For XGBoost, the main hyperparameters [8] are number of estimators (denoted as $\#estimators$), $maximum\ depth$, $minimum\ child\ weight$, $column\ samples\ per\ decision\ tree$, and $regularization\ parameter, gamma$. Hyperparameter optimization finds a list of hyperparameters and associated values that yield an optimal XGBoost model. For XGBoost models, the hyperparameter tuning was done using the *hyperopt package* which is a combination of grid search and random search approaches.

For RF, the main hyperparameters are the number of decision trees, maximum depth, and criterion for measuring the quality of the split (e.g., Gini, entropy, and log loss). Hyperparameter tuning was also completed using the *hyperopt package.*

Determining the optimal hyperparameters and structure for the DNN is more complicated than a simple grid search as used for XGBoost and RF. This requires determining the number of layers, the number of neurons in each layer, the batch size, number of epochs to train over, whether to include dropout and how much, validation split, output layer type, learning rate, and regularizers. The hyperparameters used in this research can be seen in Table 1. The number of epochs used in training varied based on validation accuracy. The DNN continued to train until the validation accuracy did not improve for five epochs in a row. A sigmoid output layer was used, so that the output was a confidence value for each classification class (healthy or waterbox fouling). Having confidence in the classification was essential for using LIME. L1 & L2 regularizers, which represent Lasso and Ridge regression, respectively, were implemented to prevent the DNN from overfitting the training data.

Table 1. DNN parameters.

Hyperparameter	Value
Number of layers	4
Nodes per layer	128,64,64,1
Batch size	64
Epochs	up to 500
Dropout	20%
Validation split	10%
Optimizer	Adam
Activation function	ReLu
Output Layer	Sigmoid
Learning rate	0.01
Loss function	Accuracy
L1 & L2 regularizer	1e-5, 1e-4

3.5 Results

The waterbox fouling prediction performance was analyzed for each model. The performance of each model was observed by introducing variations such as a missing feature or different instances of waterbox fouling. All the prediction models were further used with SHAP/LIME models to provide local explanation for each instance as well as global explanation for the entire prediction performance.

3.5.1 XGBoost Performance

To build the XGBoost binary classification model to predict waterbox fouling (unhealthy), All the CWP data except CWP 13B was considered as training data, and CWP 13B alone was considered as test data. A total of 13,768 training samples and 1,566 testing samples were used. For the binary classification model, features such as differential temperature, MOP temperature, MIB temperature, motor stator temperature, and motor current were considered. The model performance under different scenarios of data availability is discussed in the following sections. The prediction performance was observed through metrics such as accuracy, precision (number of positive class predictions that actually belong to the positive class.), recall (number of positive class predictions made out of all positive examples in the dataset), and F1-score (balances the concerns of precision and recall as $2 * \frac{precision * recall}{precision + recall}$).

3.5.1.1 All the features

Here, it is considered that all the features are available in the training data and the model is complete and comprehensive to predict waterbox fouling. For the complete data set, a training accuracy of 91.4% and test accuracy of 89.3% is achieved. The classification performance results in terms of Precision, Recall, and F-score are shown in Table 2, and they indicate the overall performance can be above 90%. A deep understanding of the influence of each feature is made possible by using the SHAP model.

Table 2. Classification performance results. (All the features are available).

	Precision	Recall	F1-score
Healthy	0.92	0.91	0.91
Unhealthy	0.91	0.92	0.91
Average	0.915	0.915	0.91

With the SHAP model, we can understand the influence of each feature on the prediction of healthy or waterbox fouling. In Figure 6, a summary plot of SHAP values for waterbox fouling associated with the feature values is shown. Considering Equation (1) with Figure 6, it is understood that higher values of DT, motor stator temperature, and motor current influence the prediction of waterbox fouling as they have positive SHAP values. Conversely, MIB temperature has the opposite trend in which low temperatures contribute to predicting waterbox fouling. MOB temperature has a mixed effect on the SHAP values: both high and low MOB temperatures are sometimes associated with waterbox fouling. This uncertainty is due to the seasonal impact on the MOB temperature measurements. Considering the overall contributions of features, DT has been identified as the most significant feature in predicting healthy and waterbox fouling (Figure 7). MOB temperature, MIB temperature, and motor current have less impact on the waterbox fouling prediction. It is also important to note that, as per plant engineers, motor current is also one of the major fault signatures. Because the motor current sometimes increases and sometime decreases with

waterbox fouling, and naturally oscillates with river temperature during waterbox fouling, the ML model is not able to characterize the impact of motor current on waterbox fouling prediction.

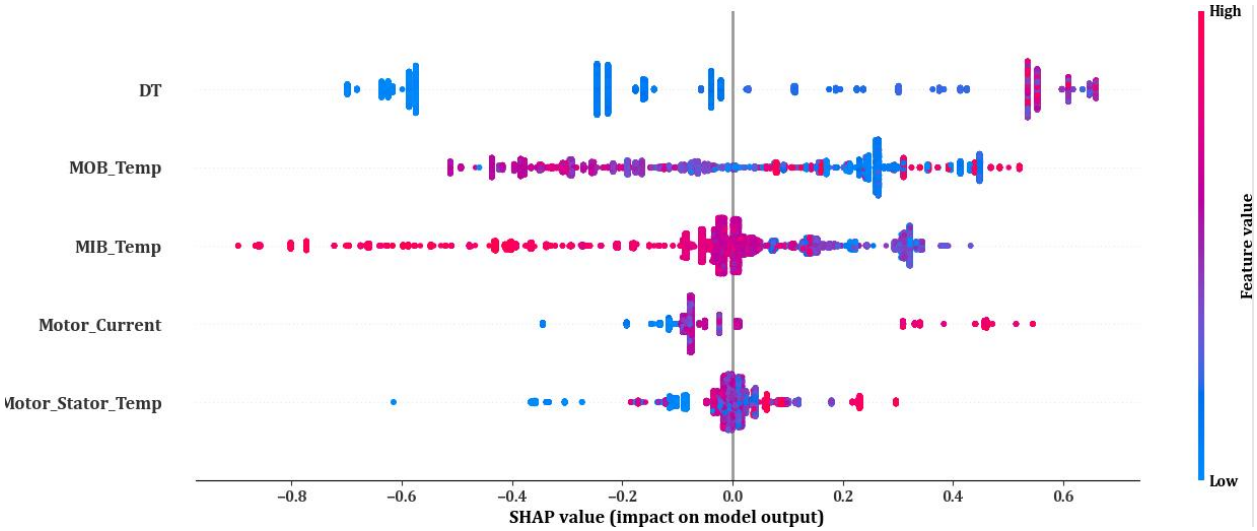

Figure 6. Feature value influence on SHAP values in prediction of Waterbox fouling.

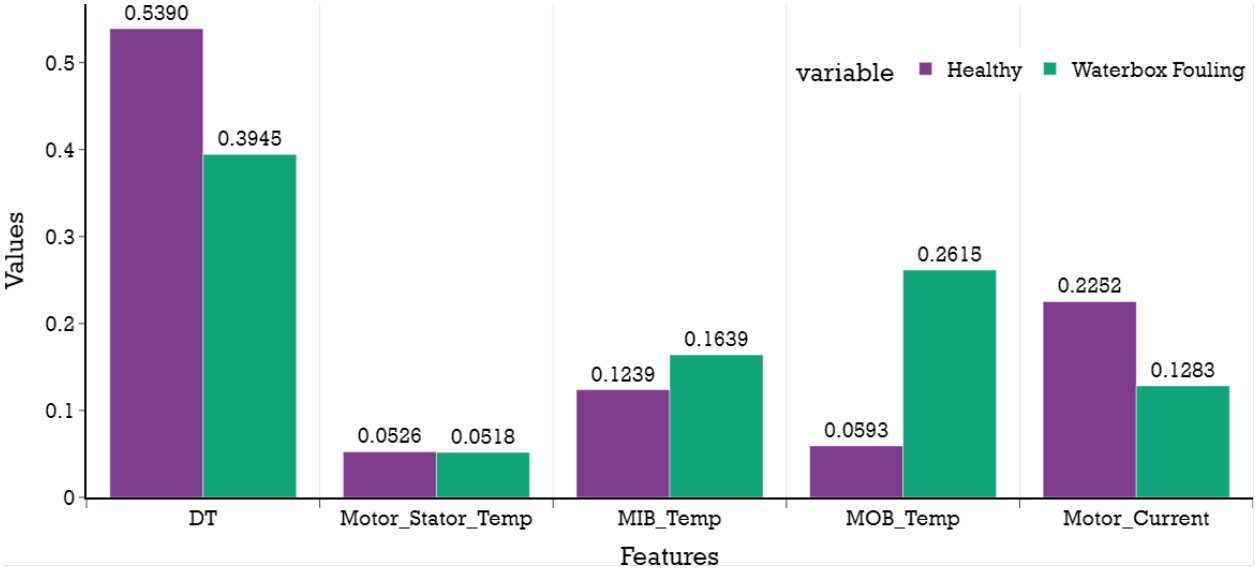

Figure 7. Feature importance for all the features in predicting waterbox fouling versus healthy data.

Figure 6 and Figure 7 provide a global interpretation (which includes all the samples/instances) method on the prediction; a local interpretation (includes a single instance/sample) method on a single sample/instance (using Equation (1)) is observed through waterfall plots in Figure 8 for the waterbox fouling prediction and Figure 9 for the healthy condition prediction. In Figure 8, the prediction is waterbox fouling since the measured DT is around 17°F and, the motor current is close to the $280 Amps$. Similarly, in Figure 9, for healthy condition prediction, the MOB temperature is above 100°F and less than the maximum operating limit. Both the motor current and the DT are within the operating limits. The color green in Figure 8 and Figure 9 means the feature at the measured instance is contributing positively to a class prediction. Whereas the color red (shown in later plots) implies the feature at the measured instance is negatively contributing to a particular class prediction. Thus, a local interpretation for an

instance can performed using SHAP values to identify the features contributing to the prediction of underlying CWP condition.

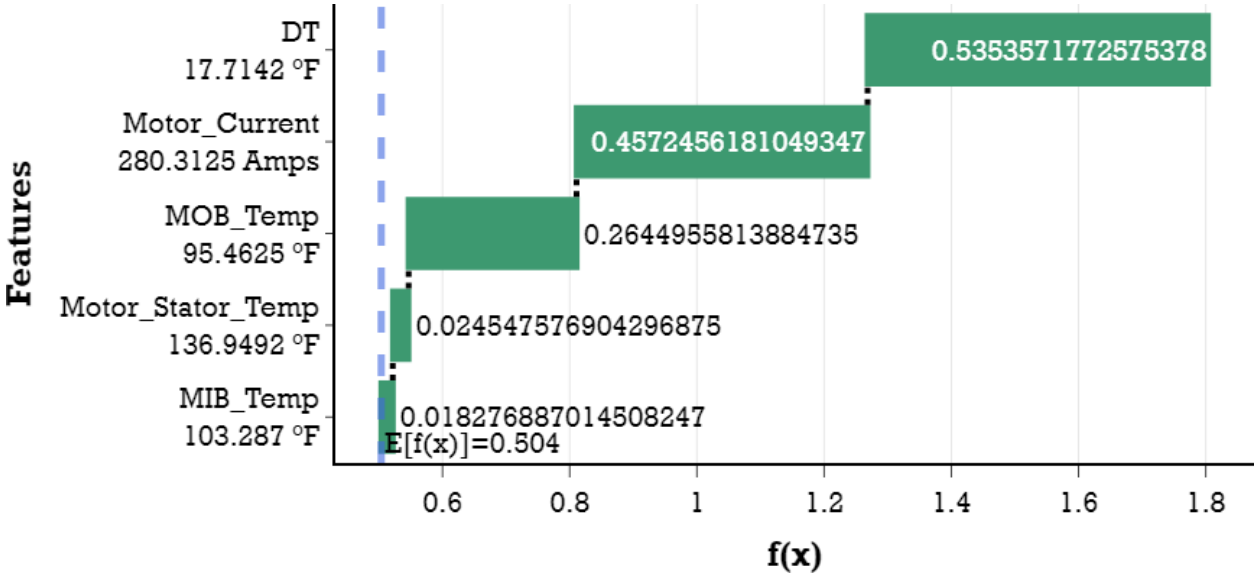

Figure 8. Local interpretation of an instance corresponding to waterbox fouling.

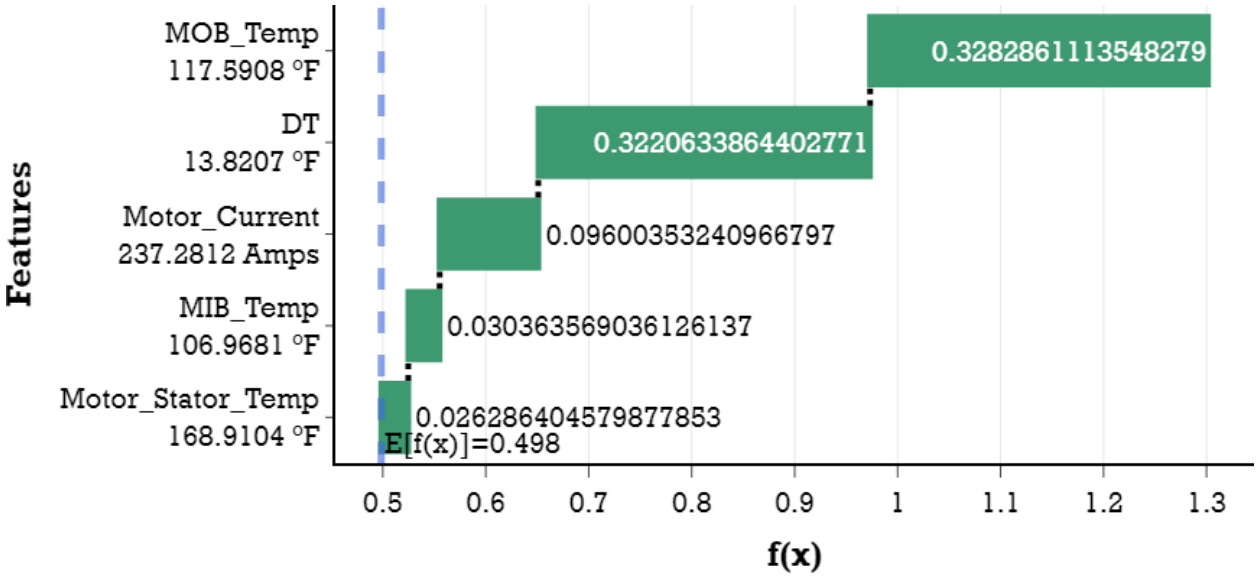

Figure 9. Local interpretation of an instance corresponding to healthy condition.

3.5.1.2 Missing feature during training

In this section, we consider scenarios of unavailability of certain measurements during training of the XGBoost model. In the first scenario, DT is missing, in the second scenario motor current is missing, and in the third scenario, MOB temperature is missing. The effect of feature values on their respective SHAP values for each of the scenarios are shown from Figure 10 to Figure 12. When DT is missing, the MOB temperatures show significant impact on larger SHAP values (both negatively and positively; see Figure 10). Essentially, low or very high MOB temperatures are associated with waterbox fouling. A similar trend can be seen when motor current is missing (see Figure 11). A higher motor current

measurement indicates waterbox fouling and low measurements indicate healthy conditions (see Figure 10 and Figure 12). The opposite behavior can be observed for MIB temperatures (Figure 10 to Figure 12). On the other hand, lower DT and motor stator temperatures indicate a healthy condition. However, the influence of MIB temperature on SHAP values is at the similar range for all the three scenarios. The feature importance and F1 score performance for each class prediction for the three scenarios of missing features during training is summarized in Table 3. The F1 scores indicate the unavailability of DT measurement could significantly reduce prediction performance while missing motor current or MOB temperatures have a marginal effect on the prediction performance.

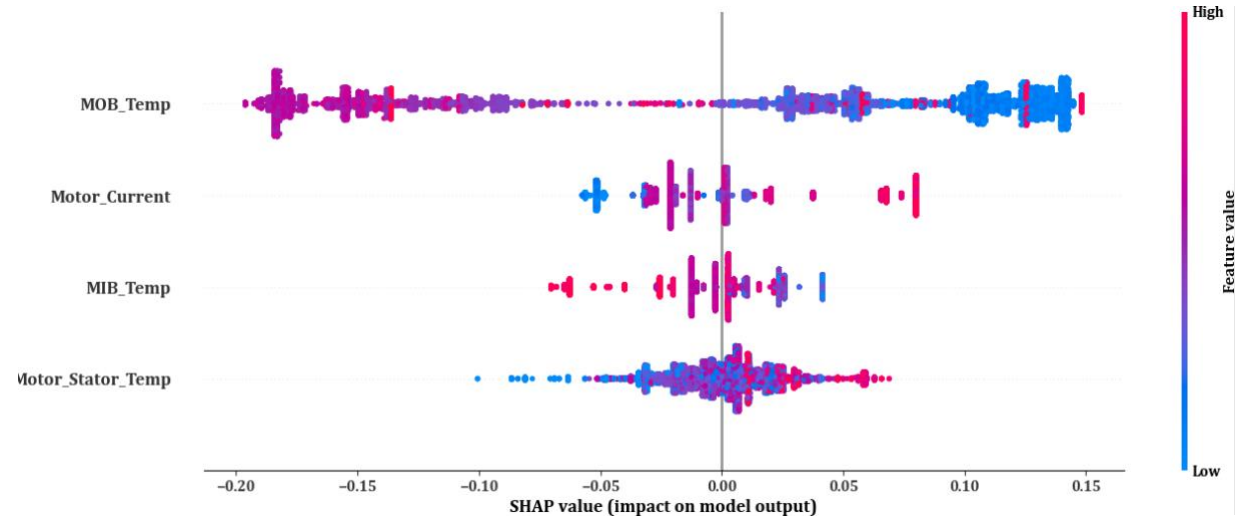

Figure 10. Feature value influence (when feature DT is missing) on SHAP values in prediction of Waterbox fouling.

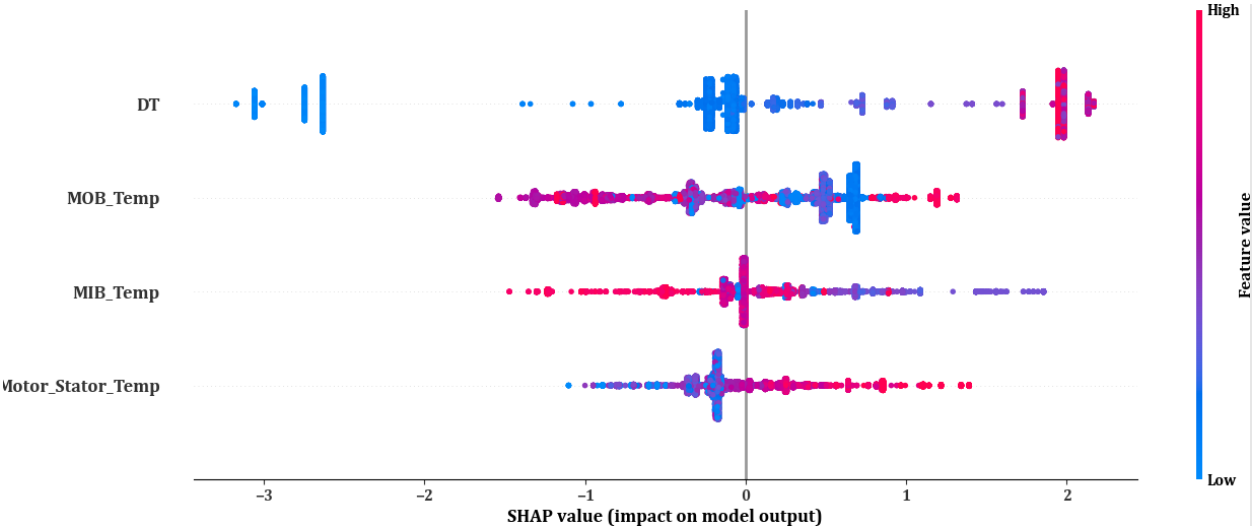

Figure 11. Feature value influence (when feature Motor Current is missing) on SHAP values in prediction of Waterbox fouling.

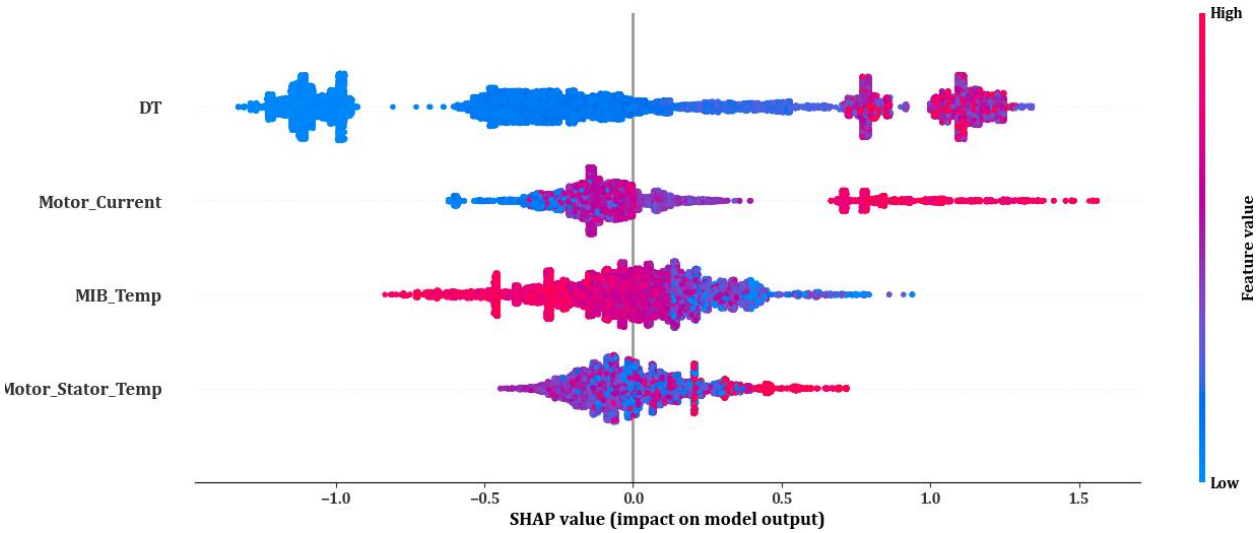

Figure 12. Feature value influence (when feature MOB temperature is missing) on SHAP values in prediction of waterbox fouling.

Table 3. Prediction performance and feature importance under unavailability of different measurements during training.

Missing Feature	Feature Importance			F1 Score	
	Feature	Healthy	Waterbox Fouling	Train	Test
DT	MOB Temp	0.490	0.637	0.83	0.88
	Motor Current	0.189	0.169		
	MIB Temp	0.172	0.100		
	Motor Stator Temp	0.150	0.094		
Motor Current	DT	0.623	0.524	0.91	0.89
	MOB Temp	0.168	0.235		
	MIB Temp	0.127	0.120		
	Motor Stator Temp	0.081	0.120		
MOB Temperature	DT	0.684	0.524	0.91	0.82
	Motor Current	0.112	0.209		
	MIB Temp	0.133	0.158		
	Motor Stator Temp	0.070	0.109		

3.5.1.3 Missing feature during inference

From the historical plant process data, it is observed that due to either technical or equipment issues, measurements might be missing. In some cases, the measurements could be missing for days to weeks. The XGBoost model enables the use of the pretrained model, which was trained with the complete feature set, even in the instances where features are missing. Note that, for the XGBoost model, when a particular feature is missing, it will be replaced as *NaN* before inputting to the model. To test prediction performance, three scenarios are considered using an instance (Date: 2018-02-02 20:00:00) from the test data. In the first scenario, no features were missing; in the second scenario, the DT feature alone is missing; in the third scenario, motor current alone is missing. With all the features available, the model

predicted the instance as waterbox fouling with $f(x) = 1.8$, which is larger than the base value, $E[f(x)] = 0.504$, with DT being the dominant feature (see Figure 13). In the second scenario, due to the missing DT feature, $f(x) = 0.57$ (see Figure 14), marginally predicting as waterbox fouling. Whereas in the third case, when the motor current is missing, $f(x) = 1.2$ (see Figure 15), predicting as waterbox fouling. Thus, when the most significant features are missing, the prediction confidence level reduces as shown in Table 4.

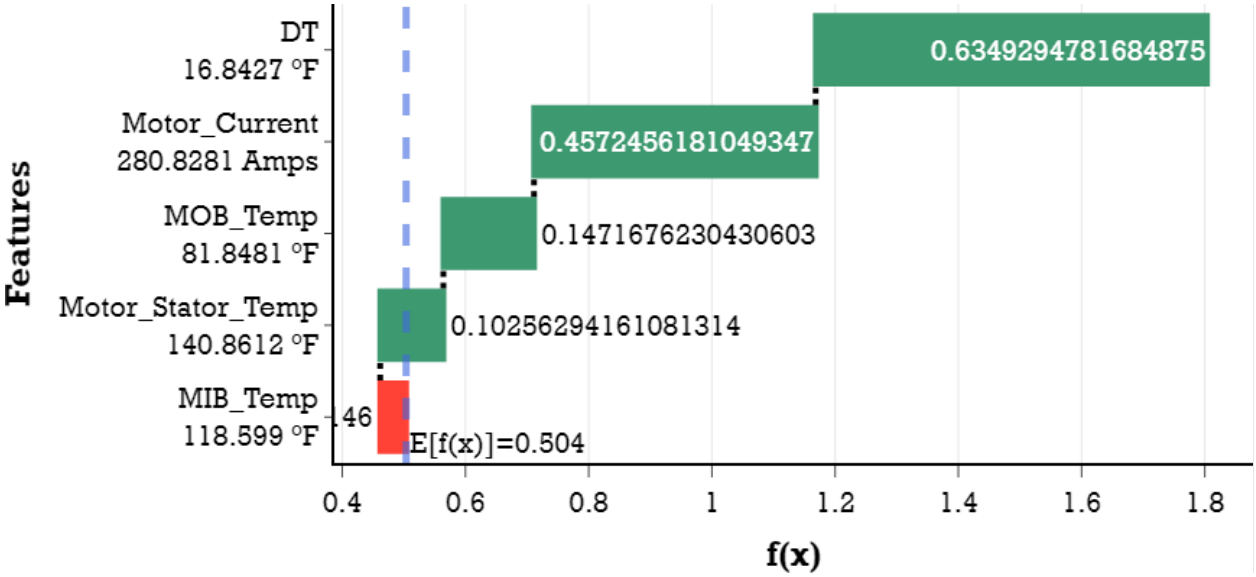

Figure 13. A test instance (2018-02-02 20:00:00) corresponding to waterbox fouling when all the features are available.

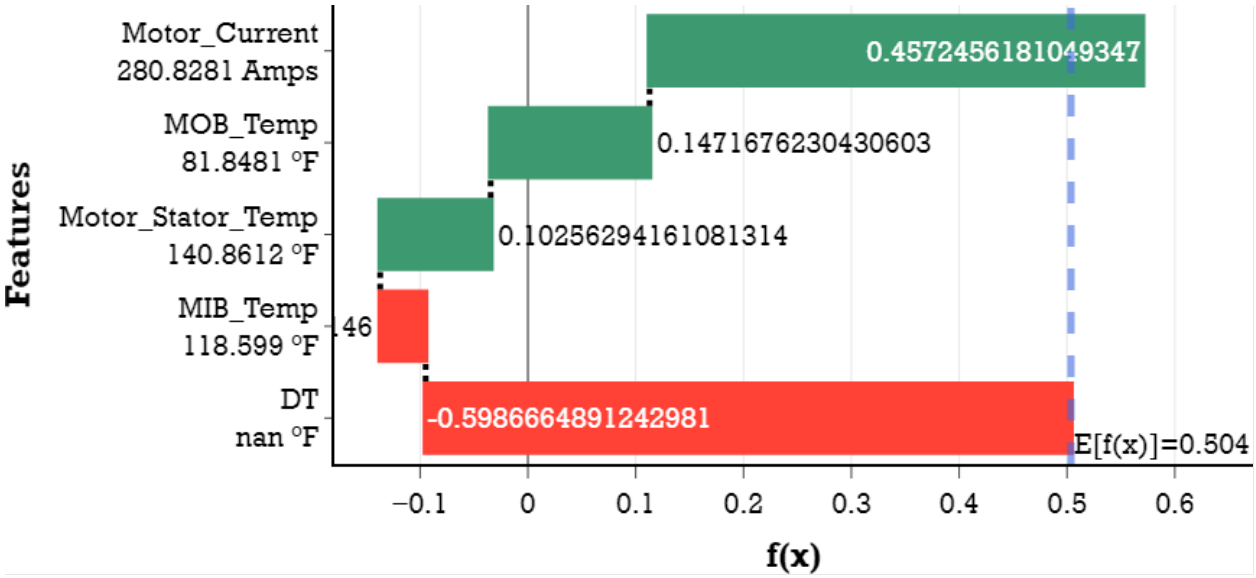

Figure 14. A test instance (2018-02-02 20:00:00) corresponding to waterbox fouling when feature DT is missing.

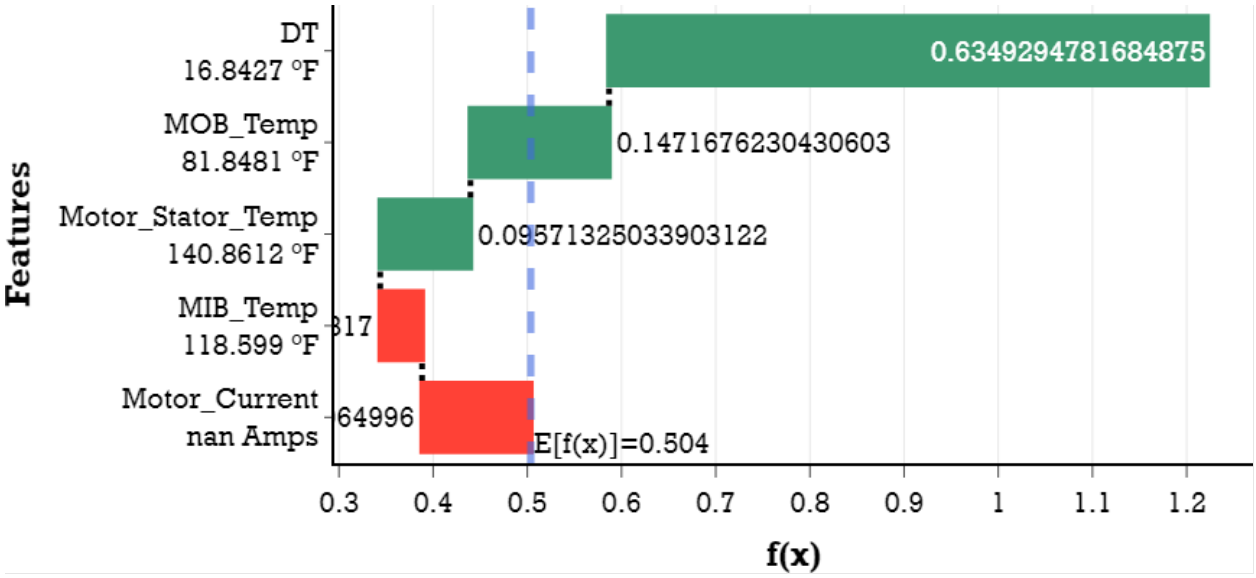

Figure 15. A test instance (2018-02-02 20:00:00) when feature Motor Current is missing.

Table 4. SHAP model prediction results for the CWP 13B instance 2018-02-02 20:00:00 for Waterbox fouling.

Missing Features	Prediction Value, f(x)	Base Value, E[f(x)]	Prediction Probability
None	1.80	0.504	0.94
DT	0.57	0.504	0.68
Motor Current	1.22	0.504	0.82

3.5.2 Random Forest and Deep Neural Network Performance

To build the RF and DNN binary classification models to predict waterbox fouling, each pump in the CWP data was considered in the training, testing, and validation data sets. The training and testing sets were divided into 4 folds, based on time, for cross validation. Figure 16 visually shows the 4 folds and the times in which they are split. The top values mark instances of waterbox fouling, while the bottom shows examples of healthy labeled data. To create the validation data set, 20% of the training dataset was separated using a stratified k-fold method. This method removes sections from training data for validation, while preserving the same percentage of each class in both data sets. Removing sections of data prevents overconfidence when training the DNN models as points will not be side-by-side in time. Also, preserving the percentage of each class in the validation set prevents the model from overfitting an over-represented class. For the binary classification models, the features DT, MOB temperature, MIB temperature, motor stator temperature, and motor current were considered. Model performance under different scenarios of data availability is discussed in the following sections.

The RF and DNN were trained and tested using cross validation. However, each fold does not contain the exact same labels in quantity, spacing, or severity. This led to differences in accuracy between each fold. Table 5 shows the accuracy of the RF when predicting a single fold. Folds 3 and 4 show unique behavior and are highlighted in yellow. Fold 3 shows a higher testing accuracy than training accuracy, which indicates that the testing data set was easier to classify or more separation was seen between the classes. For Fold 4, the RF appears to overfit the training data with a perfect training accuracy, but a less than average testing accuracy.

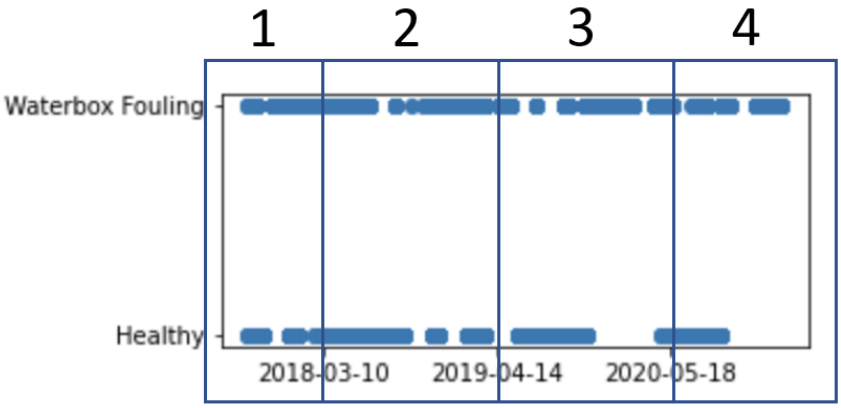

Figure 16. Training and test data were split into time segments with similar amounts of labeled data for healthy and waterbox fouling data. These time segments became folds for cross validation of the models' performance.

Table 5. RF training and testing accuracies with inputs of DT, MOB temperature, and motor current. The highlighted portion shows an instance of testing data well within the training distribution and an instance of overfitting.

Fold	RF Training Accuracy	RF Testing Accuracy
1	0.905	0.842
2	0.917	0.843
3	0.874	0.923
4	1.000	0.825

To compensate for the inherit model variability as well as the variability in the data, each fold was trained and tested 10 times for both RF and DNN algorithms. The average results and their standard deviations are shown in Table 6. Five features were available for use including DT, MOB temperature, MIB temperature, motor stator temperature, and motor current. First, all features were used as inputs. This led to the highest training accuracies but did not translate to the highest testing accuracies. Second, just the DT and motor current were used. This combination, along with inlet pressure, which was not available for this study, contains the primary features that operators currently use to determine if waterbox fouling is present, but produced below-average results as additional useful information is contained in the excluded features. DT, motor current, and MOB temperature were considered the top three most important features as determined by both SHAP and LIME, shown in Table 7. This input combination produced the most accurate models for the testing data. New information was gained by including MOB temperature, but it seems like MIB and motor stator temperatures either contain redundant or irrelevant information that did not aid the model as much in classification. From this analysis, it seemed that the MOB temperature primarily followed the ambient temperature. This can be clearly seen in Figure 17 as MOB temperature is showing a seasonal dependance, being hotter in summer months and cooler in winter months. The useful information contained in this variable may relate more to this seasonality rather than anything related to the waterbox fouling condition itself.

Table 6. Prediction performance with unavailability of different features during training.

Features	RF Training Accuracy	RF Testing Accuracy	DNN Training Accuracy	DNN Validation Accuracy	DNN Testing Accuracy
All	0.942 ± 0.045	0.838 ±0.032	0.938 ± 0.006	0.833 ± 0.049	0.826 ± 0.030
DT, motor current	0.878 ± 0.023	0.801 ± 0.033	0.882 ± 0.009	0.802 ± 0.075	0.805 ± 0.031
DT, motor current, MOB Temp	0.928 ± 0.047	0.855 ± 0.038	0.914 ± 0.008	0.826 ± 0.065	0.830 ± 0.041
No DT	0.902 ± 0.038	0.834 ± 0.039	0.907 ± 0.014	0.817 ± 0.045	0.802 ± 0.029

Table 7. Global feature importance for DNN and RF.

Feature	DNN SHAP Importance	DNN LIME Importance	RF SHAP Importance	RF LIME Importance
DT	0.216	0.291	0.219	0.276
Motor Current	0.061	0.109	0.062	0.122
MOB temperature	0.080	0.105	0.079	0.072
MIB temperature	0.050	0.089	0.022	0.031
Stator temperature	0.006	0.008	0.007	0.011

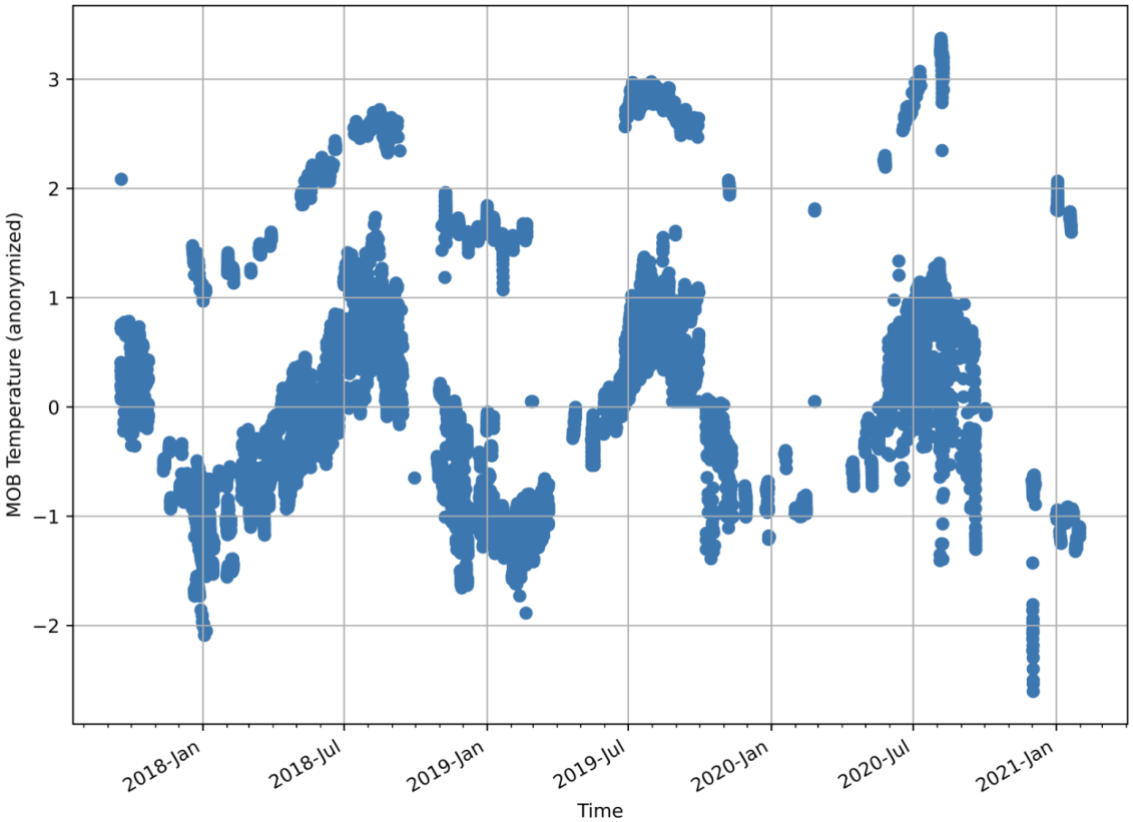

Figure 17. MOB temperature is showing a seasonal dependance as the temperatures are higher during the summer and lower during the winter.

Feature importance was determined using SHAP and LIME. For LIME, local explanations were made for every point in the training data set then the overall feature importance was the average of these explanations. Figure 18 shows an example of one of these local explanations. In this example, the RF predicted waterbox fouling with 82% confidence. The positive, green features (DT, MOB, and MIB temperature) represent a feature that supports the waterbox fouling conclusion. The red, negative features (motor current and stator temperature) do not support this conclusion. The magnitude of each feature represents how much that feature contributed to the model's prediction. In Figure 18, the DT is exceptionally high which is the primary, determining factor for the model to conclude the waterbox fouling is present. However, the motor current is within reasonable operating limits which is why that feature contributes to a healthy prediction rather than waterbox fouling. Figure 18 features some interesting insight into the MOB temperature's effect on the model. As stated earlier, this feature is related to ambient temperature and follows seasonal trends. For this local explanation, a low MOB temperature positively contributes to a waterbox fouling classification. This may be due to the increased number of waterbox fouling instances that occur in winter. This feature may be expressing that the higher likelihood of fouling is due to the season rather than any mechanical changes related to the fouling itself. This seasonal dependance on the model's predictions could be determined by using ambient temperature as an input feature.

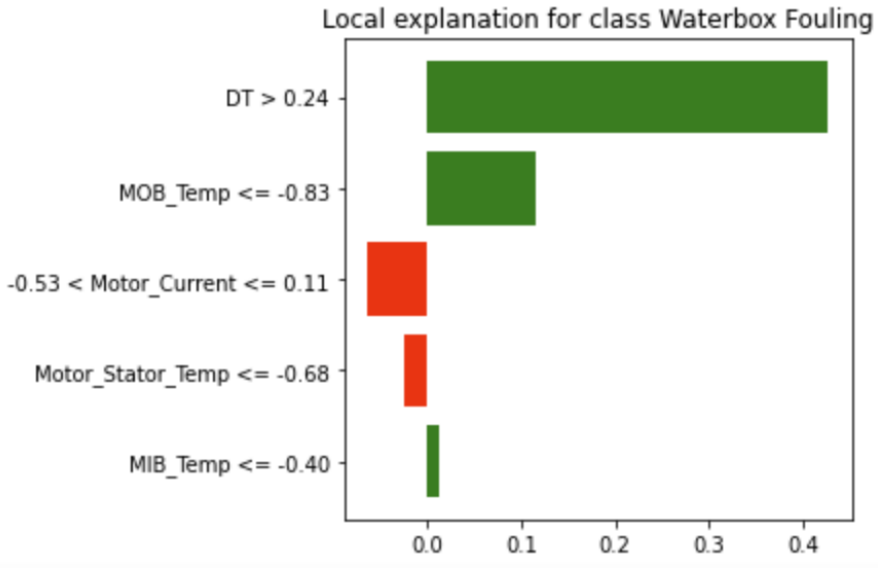

Figure 18. Local explanation of the RF's waterbox fouling prediction using LIME. Positive, green values are those contributing to waterbox fouling, while negative, red values contribute to a healthy determination. The RF has 82% confidence in this prediction. Features have been anonymized.

The overall feature importance results for RF and DNN from both SHAP and LIME can be seen in Table 7. These importance values were determined when every variable was available as input features. Overall, DT is the most important variable. This feature does not vary seasonally, and it is directly related to the performance of the pump. As waterbox fouling builds up, the DT increases. The second most important feature varies based on which metric is being used. SHAP determined that the MOB temperature was the second most important, while LIME placed higher importance on motor current. The differences seen between the algorithms are due to how they were calculated. It is noted that the best model in Table 6 used each of these top three variables. Both feature importance algorithms placed MIB temperature and stator temperature as the fourth and fifth most important features, respectively.

4. HUMAN FACTORS CONSIDERATIONS TO EVALUATE A USABLE ADOPTION OF ARTIFICIAL INTELLIGENCE BASED SOLUTIONS

4.1 Introduction

An algorithm includes any aspect of computational automation, commonly manifesting as a ML or AI model which takes input and provides an output, often in the form of a prediction or decision. In this instance, the application space for this research is the diagnosis of a nuclear reactor's cooling systems. This requires extremely complex models and interfaces to support the user in navigating the data landscape to reach a final diagnosis. The focus of this research was to capture users' mental models and attempt to match the model output to them. The model's explanation therefore needed to closely match the information that a human analyst would provide when questioned about such a prediction or decision.

This section describes the creation and evaluation of the model interface with the key considerations of explainability and trust. These are critical human factors in the adoption of automated and AI technologies. First, the key concepts of explainability and trust are introduced, followed by the literature on trust in automation generally, before discussing unique aspects to trust in automated technologies within nuclear power. Then a summary is provided of the state-of-the-art research in explainable AI (XAI) and in human-centered AI (HCAI) and the ways in which these topics—along with nuclear power's safety culture—was considered within the current research. An initial prototype model interface is then presented, which was developed to show a focused, component-level display of the ML model output in a usable and digestible form. This prototype was tested in a semi-structured interview format with M&D analysts serving as participants, and the method, results, and takeaways are summarized. The discussion section includes barriers to AI adoption, potential ways to increase trust in AI, and user mental models. Lastly, the next steps and further research related to the user-centered design and implementation of such ML modeling strategies in nuclear are presented.

4.1.1 Key Concepts: Explainability and Trust

The foundational questions described in this chapter are: "What does it mean for an algorithm to be explainable?" and "How can we develop a sense of trust in these algorithmic solutions?" A key factor in this research was trying to understand how to construct a variable of trust and how best to measure it. Trust is, fundamentally, reliance upon a person, object, process, or system [11]. Humans trust things upon which they can rely. In reference to trust in AI/ML systems, explainability is a key factor because there may be unknowns and less familiarity with the operations of the algorithmic system.

The reliance of nuclear operators on complex or arcane technologies is not new. The nuclear industry has been modernizing towards new digital systems and advanced automated technologies for several decades [12-14], and many of the engineered safety systems and their redundancies are often highly technical and can be opaque to many users. A widely used concept of automation is Sheridan's technical definition which stipulates that automation is the use of machines to change or replace duties ordinarily performed by humans [13]. This can take the form of activities performed by a machine, as well as information gathered by the machine [12]. Within nuclear operations, automation is typically employed to perform routine, repetitive tasks that may be vulnerable to human error. However, humans must trust in, and engage with, the automated systems to optimize plant safety and reliability [14].

4.1.2 Trust in Automation

Automation can eliminate human errors and greatly improve labor-intensive work systems. However, trust is one of the leading factors in whether users will rely on automated processes because failure can have significant negative implications, especially within nuclear power. Given that "automation does not simply reduce user errors but replaces them with designer errors" [15], a tension exists within emerging human-automation technologies in that designers must also work to increase trust within users.

Trust in automation has been the recipient of widespread research in recent decades [16], and from varied disciplines, each with their own definition of trust. Broadly speaking, trust in automation can be considered either as an information-based set of beliefs about the automation (cognitive), or the corresponding reliance and use of the automation (behavioral). Trust in automation technology acceptance has been shown to increase with an individual's familiarity with using the system [17].

Lee and See [18] propose a multidimensional construct of human-automation trust that contains three components:

- Performance – is the automation reliable, predictable, and competent?
- Process – how does the automation operate? Is the algorithm appropriate in fulfilling the user's goals?
- Purpose – why does the automation exist in the first place?

All three should be present for user trust to develop. When conceptualizing trust as a willingness to use the automation (i.e. resultant compliance or behavior), Lee and See [18] discuss the importance of calibrated trust (Figure 19). Calibration refers to the quality of the match between a user's trust and the system's capabilities (trustworthiness). Thus, accidents can occur when operators over-trust or under-trust automation because of failure to rely on it appropriately. The diagonal line in Figure 19 represents good calibration.

Other research suggests that trust in specific automatic controls is more predictive than trust in the overall system when it comes to operators' willingness to use the automatic functions, and that a loss of confidence in manual performance can increase the use of automation, without there being a corresponding increase in trust [19]. Additionally, user-individual differences play a role in trust in automation. For example, the age of the user is known to be a significant variable when it comes to trust in, and reliance on, automated technologies [20, 21]. For example, in many instances, older adults have been shown to have better calibration than young adults [22, 23], depending on the context. In addition, the prospective sex of an automated system's user should be considered because differences between males and females in technology communication styles have been documented [16]. Lastly, within the cognitive realm, individuals who scored high on measures of working memory showed less trust in automation than their lower working memory counterparts [24]. As the number of automated and human agents increases within networks, and interactions between the two become more complex, Miyake et al. [25] have called for investigations into team-working memory on performance, as a way to better understand collaborative capacity within team mental models.

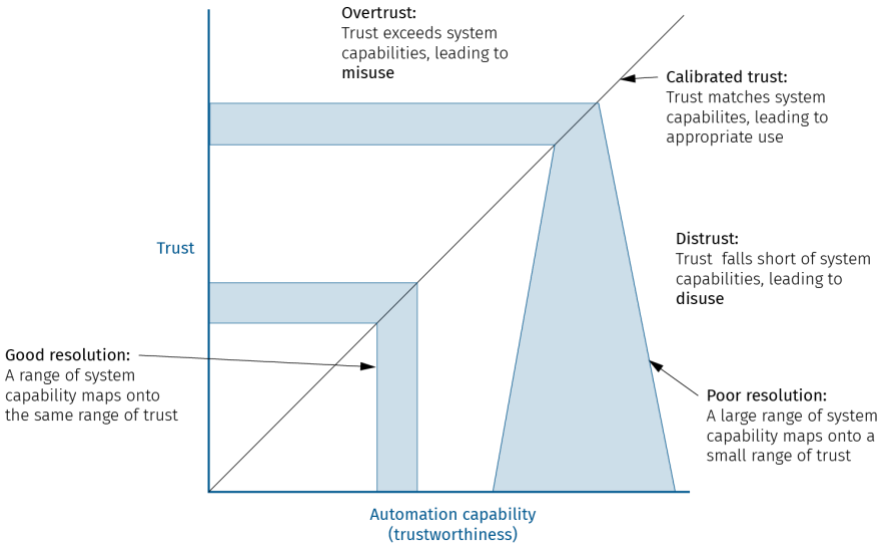

Figure 19. Relationship between calibration, resolution, and automation capability. Reproduced from Lee and See [18].

4.1.2.1 Trust in automation in nuclear operations

Some recent examples of automated modernization within nuclear operations include the Computerized Operator Support System [26] and computer-based procedures [27]. The Computerized Operator Support System is a performance aid for operators with process algorithms and fault detection that together provide an advisory system and decision recommendations. Computer-based procedures form a dynamic instruction system that incorporates digital automation of legacy, paper-based procedures. An important question within nuclear modernization is the extent to which reactor operators need to comprehend the underlying control logic and behaviors of these automated processes [28] to be able to trust and subsequently use the technology safely.

Automation transparency has been used in a bid to increase trust by making the automated functions as observable and transparent as possible to users [28]. This is important for two reasons. First, transparency aids comprehensibility, which is a key component in operators' engaging with and using the automation correctly. Second, stakeholders are responsible for ensuring the implementation of appropriate approaches to identify any faults, irrespective of the level of automation [28].

Several other trust in automation considerations that are pertinent to nuclear power are worthy of mention here. The first is perceived task-technology compatibility, or the successful integration of suitable degrees of automation to complete a task. Extensive experience with legacy technology can negatively impact the perceived compatibility of new technology, which in turn can increase mistrust [29]. The second is a lack of trust in modernization and automation upgrades being completed on time, within budget, and resulting in proven performance improvements. The third is that a distinction between trust in automation of safety versus non-safety systems has been drawn in nuclear operations, with the former historically less apt to be entrusted to automated processes. Lastly, a paradox exists within human-automation collaboration in that despite the knowledge that greater automation in industrial control brings about higher system efficiencies, operations staff are nervous that it may lead to no longer having human-in-the-loop systems.

4.1.3 Explainable AI

An evolution within automated technologies has been the introduction of AI which moves beyond pre-set and self-running programming to perform functions, and instead uses ML in which the algorithms learn from experience (feedback loops) and "understand`" the data to arrive at certain conclusions and

take actions. As with trust in automation, trust in AI has been the subject of intense research. The Ethics Guidelines for Trustworthy AI published by the European Commission [30] determined that explainability was an essential part of trust (in technology or otherwise). Explainability in AI or XAI is synonymous with transparency in automation. The logic behind XAI is that just as human decision-makers are asked to provide understandable explanations for their decisions and actions, AI must be held to the same standard. To this end, tech companies (such as Microsoft) have released guidelines for AI development that include toolkits for the issues of explainability and trust [31].

While the main goal of XAI is to increase trust in the technology, there is debate about whether the costs of understanding an algorithm can outweigh the benefits [32]. This is often termed the 'Black-Box Dilemma'—that is, we want AI to make complex decisions that are too cognitively burdensome for humans to be able to comprehend, but at the same time, we want an explanation for how the model arrived at its decisions. XAI is complicated further when it is considered that ML is typically employed in instances in which vast datasets of information are examined for complex non-linear relationships which would ordinarily be beyond human computation. Cassie Kozyrkov, Google's chief data scientist has stated that "AI does not need to be comprehensible to be trustworthy if it can be proven to perform accurately" [49]. Within process control operations specifically, there has been discussion around too much explainability leading to the system being additionally vulnerable to malicious actors (i.e., cyber-attack).

4.1.4 Human-Centered Artificial Intelligence

In contrast to XAI, there has been a recent effort to position human users more strongly in the overall AI/ML process and this perspective has been named Human-Centered Artificial Intelligence (HCAI). HCAI differs from XAI in some critical areas, but a clear difference is that while XAI can inherently be a post-hoc effort, HCAI prescribes specific design requirements for the implementation of any algorithmic solution. Particularly, that the human user be at the center of the technology's development and deployment and thus the human's requirements become model requirements.

Additionally, HCAI entails AI development with a humanistic design that supports human control [33]. Just as technologies designed with explainable AI uphold the needs of a human-in-the-loop with respect to *transparency in reaching an end-point*, HCAI serves a human-in-the-loop with respect to *controlling how to arrive at an end-point*. This sentiment moves away from historical viewpoints of AI as pure science and mathematical advancements (the rational perspective of AI), towards one that is interactive with humans and used to empower human capabilities [34]. When reaffirming that AI is in existence for the benefit of people, the European Commission HCAI expert panel has stated that AI should be designed not as an end in itself, but rather, as a promising vehicle that together increases the human potential and societal well-being [30]. To achieve this potential, [35] considers trust and cooperation between human and artificial agents as essential. Trust in AI is a central component of HCAI.

HCAI also encompasses the ethical and legal implications of AI such that technologies should be designed with an awareness that they are part of a larger system containing human stakeholders, human users, and human beneficiaries. In recent years, there has been a great deal of research and investment into issues surrounding social responsibility, accountability, and liability of intelligent systems. In combination with XAI, which promotes algorithm transparency, the HCAI framework highlights the humanistic elements behind AI by encouraging developers and designers be held accountable for their products and services [36]. Similarly, questions exist around AI technology loyalty in terms of whether it is designed to remain loyal to the developer, the user, or some other entity (e.g., the public at large). Such non-trivial issues lay at the heart of AI advancement, especially within process control industries such as nuclear power in which there must be clear cut responsibilities that can be assigned accordingly for each decision made [28]. Exacerbating this issue is the consideration that AI models evolve and adapt with new data such that future AI decisions may be unknown to the developer. In their document titled 10 CFR Part 53 Licensing and Regulation of Advanced Nuclear Reactors, the Nuclear Regulatory Commission

(NRC), the body responsible for regulating AI models used in nuclear operations, stipulates that "applicants compensate for novel designs with uncertainties." Thus, sanctioning AI technologies poses new challenges that are, to date, unresolved. These critical infrastructures merit an incredible amount of rigor and precision in any system that is relied upon.

In an interesting special issue article for the scientific journal 'Human Behavior and Emerging Technology', Reidl [37] essentially links the development of HCAI to the matter of trust in machines, given that the ways in which they solve problems are fundamentally different (or alien) to those of humans. This is because humans have evolved powerful mechanisms to understand and predict behaviors of other humans, and individuals have lifetimes of cultural experiences that help provide context for others' decisions. Machines arrive at decisions in a different manner that is unfamiliar to most. Trust increases when machines use decision-strategies that emulate our own. And as with trust in people, trust in machines decreases when we do not understand the rationale behind decisions. Indeed, a nuclear operator may be more inclined to trust ML recommendations regarding the status of the power plant, should the system not only successfully communicate the steps to reaching such a decision, but just as importantly, demonstrate that those steps match (at least to some degree) those of the operator's internal mental model should he or she have conducted the action manually.

Another conceptualization of HCAI represents a perspective change in AI design away from the either/or automation continuum, with full machine autonomy on one side and full human control on the other, towards the existence of both in parallel. This is a rethinking of the early, unidimensional formalizations of automation laid out by Sheridan and colleagues that more machine autonomy must necessitate less human control. Schneiderman [38] describes the existence of desirable systems with high levels of computer automation coupled with high levels of human control. Figure 20 describes various tasks that are better suited to either high human control, high computer automation, or both. This two-dimensional interpretation of HCAI posits that predictable and well understood tasks that require rapid deployment, such as the functions of a pacemaker or airbag, inhabit the lower right quadrant. Alternatively, complex tasks that require creative solutions such as operating an automatic camera, inhabit the top right quadrant. While decisions about exposure, focus and jitter are left to the device, the user is in control of what and how to photograph. With clever design, an optimal balance can be reached between high degrees of both automation and control. Schneiderman also cautions against the application of both excessive automation and excessive human control.

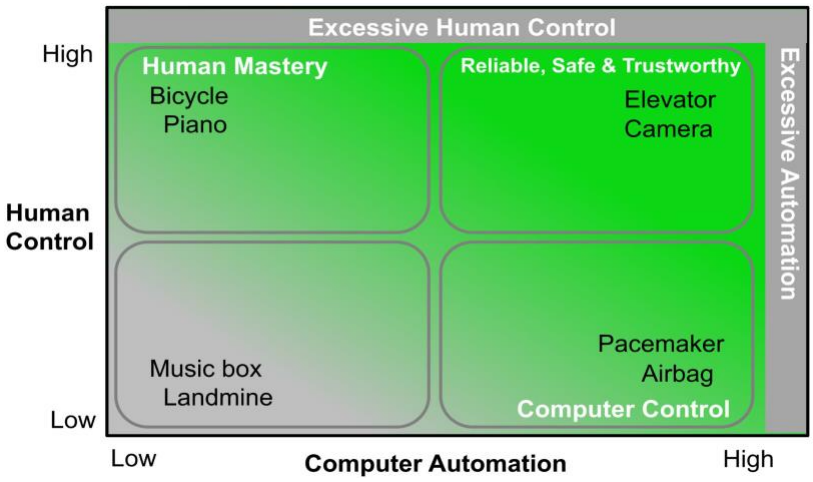

Figure 20. Examples of tasks that fall under two-dimensional HCAI.

This HCAI framework leans towards the design approach to AI (as opposed to the rationale approach). Thus, AI is not best used to replace humans, but instead used as a super tool or appliance for enhanced human performance [39]. Stemming from this notion is the recognition that computers should support social connections between humans. While AI is in existence for the good of society, machines are neither our partners nor our equals, and in the hierarchy of team membership, they are below human users. Further, Schneiderman [40] draws the distinction between trusting AI systems and delivering "trustworthy" systems. This trustworthiness comes from several external factions that represent independent oversight such as professional organizations, industry supervision, government regulators, and insurance structures. Users can decide the level of control: as trustworthiness increases, more tasks can be committed to the lower right quadrant. However, there will likely be instances in which a high degree of human control is always required, especially in the case of adverse or undesirable AI-based recommendations.

4.1.5 Nuclear Safety Culture

A challenging obstacle for trust in, and reliance on, automated and AI/ML technologies within nuclear power is the nuclear safety culture—one of the core principles that governs operations. Nuclear safety culture is a foundational part of U.S. nuclear power operations, which makes clear a commitment to the safety and protection of the public and environment over other interests [41]. In many instances, the integration of an automated or AI system increases the safety of some systems. In accordance with this safety culture practice, transparency and explainability are thus a requirement of any system with the potential to influence nuclear power operations. Across all aspects of operations, engineering, and maintenance, staff are expected to verify and validate information, data, and actions in a methodical fashion. One example of this is apparent in control room operations and three-way communication whereby an instruction must be stated, repeated back, and restated before any action can be taken.

4.1.6 Assessing Trust and Explainability in Interface

This safety philosophy, which is pervasive through all levels of nuclear power organizations, is reflected in the user research and design process undertaken for this project, and the requirement for high levels of transparency and explainability flows from this perspective. Further, a system that does not include transparency would likely fail to achieve any broad adoption.

The purpose of this research was to design an interface that communicated the results of the algorithmic decision support tool (the ML models) and tested the willingness of participants to trust the model. Specifically, the model predicted the likelihood that a waterbox fouling maintenance action should be taken. Waterbox fouling is a relatively common problem at one of the plants surveyed and requires frequent attention and maintenance due to its impact on the overall power production of the unit. M&D analysts are expected to accurately identify problems in plant systems and recommend actions to increase the efficiency of the overall plant, thus the action of recommending maintenance is a real decision that carries meaningful consequences (e.g., professional reputation at the plant). It was critical to capture a decision that would have some verisimilitude for the participants, otherwise the reported 'trust' in the system could be undermined. One challenge of this task was developing a measure of trust with strong construct validity. This was achieved by requiring participants to indicate whether they would recommend executing or delaying the maintenance action based on the information presented in the interface. We also asked the participants to use the think-aloud-procedure when using the interface, so that qualitative feedback could be captured regarding their decision-making process.

Our research did not have a strict hypothesis, however, there were some assumptions about the decisions the participants would make. For example, in scenarios of low confidence, essentially borderline cases, it was expected that participants would show hesitancy to take any action. This is reflective of the broader conservative safety philosophy at play. Beyond these expectations, there were no explicit performance measures. This effort was focused on exploring what decision the participants would make and why.

4.2 Method

4.2.1 Experimental Design

This research followed a 2 (status) x 3 (confidence level) within-subjects experimental design, depicted in Table 8, which resulted in six scenarios. The two status options were healthy and waterbox fouling. There were three confidence levels depicting the overall confidence the model had in each of the status determinations. The dependent variable representing trust in the ML model was a binary choice if users would recommend taking a maintenance action to clear the waterbox or they would delay for further information. Participants began the study from a scenario selection screen where scenarios are numbered from one to six, see Figure 21. The order of scenarios was randomized so the scenarios did not follow a pattern as the participant moved through the studies, which avoided any latent priming effects which could confound the results.

Table 8. Study design.

Status	Confidence		
Healthy	Low (50-70%)	Medium (71-90%)	High (90+%)
Waterbox Fouling	Low (50-70%)	Medium (71-90%)	High (90+%)

Figure 21. Scenario selection screen.

4.2.2 Interface Design

In most cases interfaces are designed with an inherent hierarchy and sense of depth. For example, users can often move from general information to specific information by navigating the interface, just as they can often delve into the hierarchy of data sets living behind the initial layer or landing page. For our purposes this would resemble a framework like the following: Plant-level displays → System-level displays → Component-level displays. However, it would be difficult to ensure the testing would capture only trust and not other extraneous features from the interface in such a broad set of displays. Thus, for

this effort the team focused on developing a single screen to represent a component-level display, specifically a single circulating water pump. From previous user research [1], the human factors design team understood the basic context that the analysts work from and what information they would expect from such an interface. Specifically, this includes information on the decision-making process of the algorithm and core parameters important to the CWS process, such as temperatures and motor current to contextualize the overall component performance.

The participants had stated in previous research that the potential for decreased power production, and thus revenue, was an important factor in evaluating the possibility of taking a maintenance action as the cost of maintenance actions can be significant. This is reflected in the gross load representation within the interface. Table 9 shows each designed feature, definition, and the corresponding location of the feature on the interface in Figure 22. The goal of the interface design was to give the participants a minimal baseline of information to make the prescribed decision to attempt to isolate the primary variable of trust in the model. Most of the information presented to the user was from the model directly, and the minimalistic design strategy attempted to isolate the user's decisions from even their own diagnostic ability, hence no further data or "drill down" features are present. The participants were to make the decision based only on the information given to them.

As shown in Figure 22, the status and confidence levels are reported in area D. The choice was made to not immediately display probabilities to the participants. This decision was based on best practices regarding human psychology and our poor inherent statistical understanding [42]. The representation of raw probabilities to users can be problematic as users often overestimate positive results and underestimate negative results when presented with numerical probabilities. As humans are not natural statisticians [42] it is best practice to avoid any design which rely on sound statistical reasoning. However, the confidence percentages were not absent from the design. For each scenario this information was accessible by hovering over the confidence determination, (i.e., the text 'Low' in Figure 22).

Table 9. Interface features and locations.

Feature	Definition	Location
ML model decision (status)	Does the model think the pump is healthy or experiencing waterbox fouling?	D
Confidence in decision	The confidence the model has in the above determination.	D
Parameters of interest	For this study these parameters were determined to be 'of interest' by participants.	A
Metrics for feature importance	These are the metrics which the model judged the normality of the parameter in question.	B
Feature importance ranking graph	This graph corresponds with the parameter metrics next to it and shows the order of feature importance with a visual to demonstrate magnitude differences in importance.	C
Economic context	As the lost revenue, 'derate %,' is a critical factor in the participants' decision-making it is represented here.	E

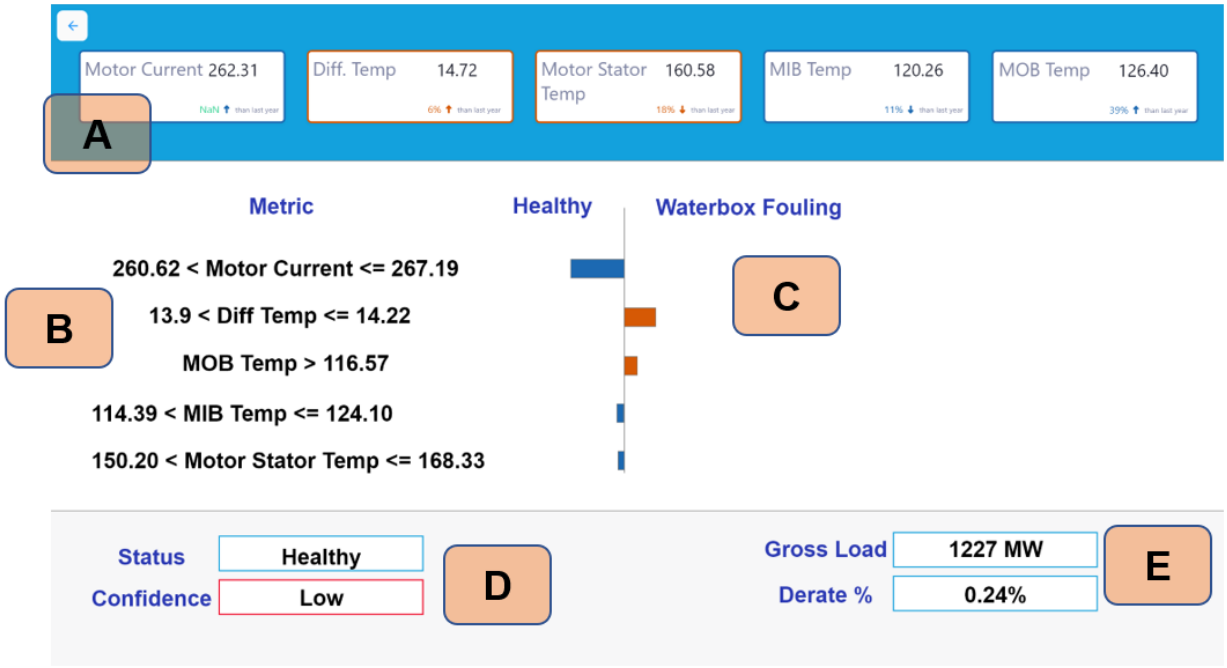

Figure 22. Interface with location markers.

4.2.3 Procedure

Two participants were each presented the six scenarios with alternating conditions of model prediction (healthy or waterbox fouling) and model confidence (low, medium, or high). Participants employed a think-aloud protocol as the participants were shown the interface and the experimenter controlled the mouse (this was due to technical difficulties in sharing the prototype directly) with minimal engagement by human factors staff. They were asked to verbally acknowledge the content of the interface they were interested in at any given time during each scenario (e.g., hovering over the confidence level to reveal percentage), before verbalizing their maintenance action decision based on information provided in the model for each scenario. If participants did not freely elaborate on what factors led to their decision, they were explicitly asked. After all scenarios were completed, the participants provided verbal feedback on the model, the interface design, features, components of the interface and anything else they noticed throughout the study. Responses were recorded manually and analyzed after the study ended.

4.3 Results

The results are divided into two parts. The first section covers performance on the maintenance decision tasks with a discussion on specific explainability feedback the participants raised. The second section covers feedback on the interface design and how design choices supported explainability as well as recommended changes. In keeping with best practices in the field, this mixture of quantitative and qualitative data has together been shown to produce meaningful findings for usability studies.

4.3.1 Results of Maintenance Decision Task

The results for each scenario are shown in Table 10. For the four trials with stated expectations, User A matched twice, and User B matched three times. In both instances and for both users, performance matched expectations with the decision to delay maintenance action. The participants only elected to trust the model and take a maintenance action once and showed deviation from the expected response three times for User A and twice for User B. The ‘None’ expectations reflected the low confidence levels of each status decision, and it was not clear how participants would respond to those borderline cases.

Table 10. Results of decision task.

Scenario	Status	Confidence	Participant A decision	Participant B decision	Expectation
1	Healthy	Medium	Delay	Delay	Delay
2	Waterbox fouling	Low	Delay	Delay	None
3	Healthy	High	Delay	Delay	Delay
4	Waterbox fouling	Medium	Delay	Delay	Action
5	Healthy	Low	Delay	Delay	None
6	Waterbox fouling	High	Delay	Action	Action

In understanding why the participants deviated from expected decisions and elected not to trust the model, even in high confidence situations, the qualitative debriefs were valuable. Two factors had a large impact on the participants' conservative decision to delay action in favor of further investigation. First, the model included three parameters that were not viewed to have diagnostic or predictive value to the participants, namely motor stator temperature, MIB Temperature, and MOB temperature. In scenarios where these parameters ranked highly in the feature importance for the model, their inclusion and high ranking led to more skepticism in the model and a stated need for more investigation. Secondly, and relatedly, the participants expected any model to largely mimic their own mental models and diagnostic processes. This sentiment was shared by both participants as they indicated that if the model largely made decisions that they would also arrive at a large majority of the time they would trust it more. If it deviated and made a decision that, in their experience, was not supported by the data then it led to decreased trust in the model. This is an important result for future model development as matching user mental models isn't often the focus of those tasks. A specific interface change that applies here is the requirement to have trending capabilities within the interface. Both participants stated that being able to view the parameters across various time contexts would be helpful in increasing trust in the model. In conclusion, it appears that participants were not keen on trusting the model in nearly all cases.

One of the discussed requirements for this study was the level of transparency. From previous user research the design team knew that the participants would expect to be able to see all parts of the model's decision process. In general, this aspect of the design received positive feedback. Participants were able to see and understand what the model had done and how it arrived at its decision. This was not explicitly measured with a dependent variable but was a focus of the debriefing after the initial run through of scenarios. It is also important to note that there are some issues with the construct of trust in this experiment. Two features were intended to underlie trust and provide some measures for evaluation, namely whether the participant agreed with the model's determination and if they selected the option which the experimenters had expected. However, neither truly held up against a construct of trust. The model did not give a recommended action to the participant to agree to and in many cases the participant did choose what we expected but with no realistic feedback on if that was correct or not, which is difficult to evaluate. These lessons will be reflected in future work.

4.3.2 Interface Design and Model Feedback

Each participant was asked to evaluate the interface design based on scenario conditions, especially pertaining to the predictive models. Participants were asked to provide this feedback aloud throughout the scenarios as well as post-study. Throughout the study, both participants expressed resistance to trusting the models. Neither participant trusted the model as a single source of data, and when a decision was made where the model was considered a factor, it was within scenarios that were obviously and easily verifiable at-a-glance of "healthy" or "waterbox fouling" (e.g., scenario 3, "healthy, high"). In other

words, participants only "trusted" models within scenarios that were quickly identifiable and did not require multiple minutes of manually verifying the model.

Participants also seemed to conflate "trust" with the ability to verify the model quickly and manually. Participant comments like, "I would trust this model more if trending history was provided for motor current in the same display," demonstrates a lack of trust in the model at face value. Researchers expected participants to cross-verify the model with the explainability data included in the interface, however, seeking additional information beyond what was included in the interface to validate the model accuracy suggests an overall lack of trust in the model. When participants were asked about their lack of trust in the models, each participant suggested including additional information in the design. They made recommendations such as providing supplementary data that could be used to validate the model accuracy.

General feedback surrounding the content of the interface and the model data were also evaluated. The main insights can be summarized as follows:

- Both participants seemed surprised that MIB and MOB temperatures were a part of the model. Each stated that MIB and MOB temperatures are not useful parameters on their own to assist with a diagnosis of waterbox fouling. Participants recommended removing MIB and MOB from the main parameters displayed in the model.
- Participant A stated that a more appropriate factor of waterbox fouling is the overall expected gross load and any potential deviation from that and recommended including this in the model and design.
- Both participants described scenarios where providing the additional context of data would help them "trust" the model more. For example, providing the expected motor current rate in addition to the current output would help them more intrinsically understand how likely waterbox fouling is. Both participants recommended adding expected motor current rate alongside the current output to the design.
- Both participants also stated that providing additional at-a-glance trending information would lead to quick manual verification instead of having to click to access history which would help them make a decision more confidently and quickly. Both participants recommended adding additional trending information into the design.

Each participant's feedback regarding the design of the interface can be summarized as including additional data that is intrinsically meaningful to operators to allow further opportunities to quickly validate the predictive models included in the interface. Therefore, at this stage, it still appears that the model, as a single source of decision support, would not be sufficient. Additionally, the absence of trend data was not an intentional feature, but the trend data was left off due to constraints within the scenario outputs from the model that prevented incorporating trend data for the parameters. Future design iterations should consider including these features.

4.4 Discussion

This section explored some aspects of explainability, trust, and acceptance of algorithmic solutions like AI/ML models in the nuclear power industry. Acceptance of an algorithmic solution in the decision-making process of different tasks within the nuclear power industry can be challenging. The focus of this research was asking what would obstruct or assist the acceptance of such models in a maintenance and diagnostic focused organization and as such, it was necessary to truly explore what participants are expecting in terms of the overall interrogability and transparency of these models to justify accepting the model's determinations. Another particular challenge is the reality of nuclear safety culture and what that means for the development and implementation of such a system. The specific perspective of construction, transparency, and explainability that participants have is nearly directly tied to the principles underlying the nuclear safety culture. Next, the design and user testing process of highlighting the

explainability problem were explored. An interface was developed specifically to present scenarios to the participants which highlighted the model determinations and how it arrived there. The interface was populated by real model data and plant parameters which helped establish the proper and familiar context the participants would expect. The results of the testing showed conclusively that the participants did not exhibit any significant trust in the model's determinations unless the model matched their own conception of the scenario which introduces doubt into those scenarios where their decision matched the expected result. If the participants were predominantly using their own cognition as the measuring stick, then were they trusting the model or their own minds? This section explores further takeaways from this research from the human factors team.

4.4.1 Potential Barriers to AI/ML Adoption in the Nuclear Industry

There are two main barriers to predictive model integration into nuclear industry operations: the cultural issue of a general lack of trust in single sources of information and the overall efficacy of predictive models in a risk-averse industry.

Lack of Trust

There is a cultural framework in the United States nuclear industry of manually verifying multiple sources of information in a power plant before making a decision. Even when nuclear operators have high confidence in a data value source, they will verify that value with additional sources of information before taking an action. This culture is reinforced through procedure-centric operations and three-way communication. This makes it difficult to incorporate predictive models that operators readily rely on into everyday operations. Since a primary advantage of predictive model integration into nuclear operations is to reduce the amount of time required to make an operations decision, this advantage is only possible if operators trust the predictive models. If operators cannot trust the models, they will spend just as much time if not more to making operations decisions as the models will just be another data point to validate their plant status consensus. To truly combat this cultural framework, predictive models must incorporate an operator's mental model and incorporate the needed data values and present them in an intrinsically meaningful way.

Predictive Model Efficacy in the Nuclear Industry

Predictive models can aggregate data and present it to an operator to reduce workload and overall decision time. However, predictive models are typically used in low-risk, industrial operations. Nuclear operations are risk-averse and very safety-centric. As a result, there is hesitancy to incorporate predictive models into nuclear operations as model accuracy needs to be perfect. An additional concern is, if a model makes a mistake, and an operator relies on that model, who or what is at fault? Regardless of model accuracy, if there's even a slight chance that the model could be wrong, operators will not trust it to make decisions. Creating models with 100% accuracy is directly dependent on the availability of data that is used to supply the models. As additional methods of capturing and consolidating data emerge, the idea of developing predictive models with perfect accuracy is becoming more attainable. However, considering the current technological capabilities of capturing data in the nuclear industry, the perfect accuracy of AI/ML models is more conceptual than concrete.

4.4.2 Increasing Trust in AI

XAI and HCAI are predicated around increasing trust, and by extension, the use of the technology. Essential to this worldview is the recognition that algorithms can produce mistakes. ML is a technology that is only as reliable as the data it is given, which can be biased. An interesting paradox exists in that while humans are expected to be flawed, often automation is expected to be fail safe [43]. However, the data, and the models produced from them, are imperfect. Research has shown that while initial trust with unfamiliar systems may be high (likely a result of the belief that the automation contract always be upheld), trust plummets after an interaction involving an automation error, to an unwarranted degree [44]. This can be mitigated somewhat after the fact by providing explanations as to why the system might err.

It is possible then that while an upfront disclaimer of a system's potential for mistakes may initially discourage trust, should a mistake occur, provided there is transparency as to why, trust will not be lost entirely.

In other research, Merritt [45] demonstrated that perceptions of machine characteristics accounted for over half the trust variance in human-automation interactions above actual machine characteristics. This finding speaks to the salience of user expectations in the likelihood of errors. Moreover, transparency in these failures to participants is key to establishing and maintaining trust. As with human relationships, research points to intelligent agents apologizing after errors to help repair trust [46]. Finally, the importance of gathering failure feedback and reporting near misses has been stressed in maintaining trust in AI. These data can then be analyzed to support PdM and increase reliability [38].

4.4.3 Matching User Mental Models

An important finding from this research was the direct statement that the participants would trust the model if it made decisions as they do. It isn't a surprising takeaway; it is common that people would trust something like them. However, in this instance, the participants weren't evaluating trusting another person or some other socially grounded construct, which is where this commonality bias commonly comes into play. Instead, the participants were evaluating their own reliance on a technical, computational feature and still expected a similarity in decision-making. The main takeaway for future work is that AI/ML model developers should work closely with human factors teams and participants to ensure that the models are built in a way which aligns well with user mental models and therefore builds trust in the system. This angle doesn't require complete adherence with a user's mental model. If a model simply replicates a human analyst, then it isn't an effective tool, as computational methods and models can provide significant performance improvements to traditional human cognition and decision-making. Rather, there is a need to closely align with the way participants make decisions so that participants can generalize their own expertise to the model and then novel or unique decisions the model makes will be more likely to be received well by the participants.

4.5 Potential Research Gaps

This section explored concepts of explainability and trust, the nuclear safety culture, and a small-scale study focused on these concepts to evaluate the particular expectations of users in trusting a model-based decision support tool. While the results showed that the participants tended to not trust the model's recommendations without further information and data (such as trends), important insights were gleaned about why they wouldn't trust it and what would help them trust it more. This section presents several possible angles of future research which can better capture the nuances of the diagnostic processes the participants follow and how the model should be constructed to best support that.

4.5.1 Explainability and Trust Construction

The most immediate next step in this research path is to better understand the constructs of explainability and trust. A better characterization for what a trusting action is, how explainability supports trust, and how those concepts are evaluated in developed interfaces. There is a strong need for robust constructions of these variables for better experimental research to occur, particularly in the nuclear power field. This research took an initial pass but fell short of capturing and evaluating a clear case of trust in several critical ways. The experimenters look forward to the next iteration of these concepts and better development for future research.

4.5.2 Deeper Interface Development to Support Diagnostic Processes

An important piece of feedback received was the need for more context around the data presented, particularly a graphing or trending representation, and the expectation of the model mimicking the participant's mental model. The interface developed in this research was limited to static data representations, so it lacked any dynamic or time-based representation of the represented parameters.

Future work should prioritize developing a more fully-fledged interface to support exploration of data, trends, or other features in addition to the model-based information shown in this design. Supporting the existing mental models with the addition of a ML model would help human factors researchers to isolate when users are trusting the model and when they are performing their own diagnostics.

4.5.3 Matching Model Process to User's Mental Models

Participants stated explicitly that they would be more inclined to trust the model if it made decisions the way that they do, and therefore they would agree with not just the decision but the path the model took to get there. Future work should be done in understanding how to translate user requirements and mental models into inputs to the ML model development process. If ML models more closely reflect the decision processes of the users they are supporting, then significant gains in trust could be possible.

4.5.4 Nuclear Safety Culture Challenges

Another key area of future work should be in better understanding of how these model-based systems fit within the current nuclear safety culture and its expectations. It appears that the more opaque a model is the greater the conflict it will have with these fundamental principles. As nuclear human factors are extremely grounded in implementation regulations and verification & validation steps, this is a relatively immediate concern of the human factors teams. This angle should focus on the broader expectation and philosophy of how these automated models can exist in an environment that places high rigor on validity of methods and systems.

5. SUMMARY AND PATH FORWARD

In this report an initial technical basis for developing explainable and trustable AI and ML technologies was presented using the forward-backward closed-loop process. The technical basis was evaluated for a specific case study on waterbox fouling problems in the CWS at a power plant. The performance of ML models developed was evaluated for different data variabilities and the variabilities were explained using an objective metric like LIME. A user-centric visualization was developed to verify user acceptability of the ML outcomes for six different scenarios related to the waterbox fouling case study. The work reported in this report lays the foundation for addressing the adoption challenges of AI/ML technologies across plant assets and the nuclear fleet to achieve risk-informed PdM strategies at commercial NPPs.

The technical basis developed in this report will be extended to other fault modes and to a wide user verifiability study. For the next year, the research scope will focus on the verification and validation of the explainability and trustworthiness of AI/ML technologies. As part of the path forward, another feature that may aid in the trustability of ML is novelty detection, which will be explored in detail. Novelty detection is the task of recognizing differences between a test data point and the training data distribution [47]. This is important, because in regression tasks, ML typically performs better with interpolation rather than extrapolation. Interpolation is when the test data point lies within the training distribution, while extrapolation is when the test data point is outside this distribution. In Figure 23, several example models are fit to training points within the white area. Within this distribution, the models accurately predict the training data. However, once outside the training distribution, in the gray area, these models begin to diverge and model performance is no longer guaranteed. With the use of novelty detection, test data points that are not within the training distribution can be found and flagged to indicate that model performance on this data point may be inaccurate. Several novelty detection algorithms currently exist, such as Local Outlier Factor and Isolation Trees, but their usefulness in improving ML trustability needs to be thoroughly tested before use.

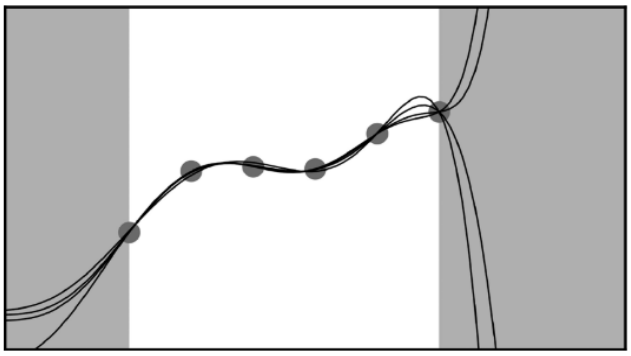

Figure 23. Graphical representation of a model being fit within the training distribution (white area) and how the extrapolation into new domains (gray area) may vary. Figure from [48].

6. REFERENCES

1. Agarwal, V., et al. (2021). Scalable Technologies Achieving Risk-informed Condition-Based Predictive Maintenance Enhancing the Economic Performance of Operating Nuclear Power Plants," Idaho National Laboratory, INL/EXT-21-64168, Rev. 0.
2. Agarwal, V., et al. (2021). Machine Learning and Economic Models to Enable Risk-Informed Condition Based Maintenance of a Nuclear Plant Asset," Idaho National Laboratory, INL/EXT-21-61984, Rev. 0.
3. Agarwal, V., et al. (2019) "Deployable Predictive Maintenance Strategy Based on Models Developed to Monitor Circulating Water System at the Salem Nuclear Power Plant," Idaho National Laboratory, INL/LTD-19-55637, Rev.0.
4. Idaho National Laboratory. 2020. "Light Water Reactor Sustainability Program, Integrated Program Plan," INL/EXT-11-23452, Rev. 8.
5. F. Poursabzi-Sangdeh, et al. 2018. Manipulating and measuring model interpretability. arXiv preprint arXiv:1802.07810.
6. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 2017-Decem(Section 2), 4766–4775.
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. https://doi.org/10.18653/v1/n16-3020.
8. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.
9. Chornovol, O., Kondratenko, G., Sidenko, I., & Kondratenko, Y. (2020). Intelligent forecasting system for NPP's energy production. *Proceedings of the 2020 IEEE 3rd International Conference on Data Stream Mining and Processing, DSMP 2020*, 102–107. https://doi.org/10.1109/DSMP47368.2020.9204275.
10. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011.
11. Sullins, J. P. (2020). Trust in robots. In The Routledge handbook of trust and philosophy (pp. 313-325). Routledge.

12. Boring, R. L., Ulrich, T. A., & Mortenson, T. J. (2019). Level-of-automation considerations for advanced reactor control rooms. Proc. 11th Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, Orlando, Florida, February 9–14, 2019, p. 1210, American Nuclear Society.
13. Lin, C. J., Yenn, T.C., & Yang, C.W. (2010). Automation design in advanced control rooms of the modernized nuclear power plants. Safety Science, 48, 1, 63. https://doi.org/10.1016/j.ssci.2009.05.005.35.
14. Skjerve, A.B.M., & Skraaning Jr., G. (2004). The quality of human-automation cooperation in human-system interface for nuclear power plants. International Journal of Human Computing Studies, 61, 5, 649 https:// doi.org/10.1016/j.ijhcs.2004.06.001.37.
15. Kima, Y., & Park, J. (2018, October). Envisioning human-automation interactions for responding emergency situations of NPPS: A viewpoint from human-computer interaction. In *Proc. Trans. Korean Nucl. Soc. Autumn Meeting*.
16. Hinze, H. (2022, January 27). Peeking into the Black Box – Trust in AI – Part 2. *Wolters Kluwer. https://www.wolterskluwer.com/en/expert-insights/peeking-into-the-black-box-trust-in-ai-part-2.*
17. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.
18. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392.
19. Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243-1270.
20. Czaja, S. & Sharit, J. (1998). Age differences in attitudes toward computers. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences. 53.* P329-40. 10.1093/geronb/53B.5.P329.
21. Ho, G. & Kiff, L., Plocher, T., & Haigh, K. (2005). A model of trust and reliance of automation technology for older users. *AAAI Fall Symposium - Technical Report.* 45-50.
22. Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*, 1059–1072.
23. Sanchez, J., Fisk, A. D., & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. *In Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting (pp. 586–589)*. Santa Monica, CA: Human Factors and Ergonomics Society.
24. Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science, 18 (6)*, 573 -591.
25. Miyake, T., Parker, C., Hall, A., & Boring, R. (2021, September). Conceptualizing Team Working Memory: Implications for Human-Automation Collaboration. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 65, No. 1, pp. 1259-1263). Sage CA: Los Angeles, CA: SAGE Publications.
26. Boring, R. L., Thomas, K. D., Ulrich, T. A., & Lew, R. T. (2015). Computerized operator support systems to aid decision making in nuclear power plants. Procedia Manufacturing, 3, 5261-5268.
27. Oxstrand, J., & LeBlanc, K. (2014). Computer-based procedure for field activities: Results from three evaluations at nuclear power plants (No. INL/EXT-14-33212). Idaho National Lab.(INL), Idaho Falls, ID (United States).

28. Skraaning Jr, G., Jamieson, G. A., & Joe, J. C. (2020). Towards a Deeper Understanding of Automation Transparency in the Operation of Nuclear Plants (No. INL/EXT-20-59469-Rev000). Idaho National Lab.(INL), Idaho Falls, ID (United States).

29. Kovesdi, C. (2021, July). Examining the Use of the Technology Acceptance Model for Adoption of Advanced Digital Technologies in Nuclear Power Plants. In *International Conference on Applied Human Factors and Ergonomics (pp. 502-509)*. Springer, Cham. INL/CON-20-61075-Revision-0.

30. Ethics Guidelines for Trustworthy AI (European Commission, 2019); https://ec.europa.eu/digital-single-market/en/news/ethicsguidelines-trustworthy-ai.

31. Liao, Q. V., Pribić, M., Han, J., Miller, S., & Sow, D. (2021). Question-driven design process for explainable AI user experiences. *arXiv preprint arXiv:2104.03483*.

32. Bunt, A., Lount, M., & Lauzon, C. (2012, February). Are explanations always important? A study of deployed, low-cost intelligent interactive systems. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (pp. 169-178).

33. Xu, W., & Dainoff, M. (2021). Enabling human-centered AI: A new junction and shared journey. *arXiv preprint arXiv:2111.08460*.

34. Auernhammer, J. (2020) Human-centered AI: The role of human-centered design research in the development of AI, in Boess, S., Cheung, M. and Cain, R. (eds.), *Synergy - DRS International Conference 2020*, 11-14 August, Held online. https://doi.org/10.21606/drs.2020.282.

35. Kuipers, B. (2022). Trust and Cooperation. Front. Robot. AI 9: 676767. doi: 10.3389/frobt. 2022.676767 Trust and Cooperation. *Responsible Robotics: Identifying and Addressing Issues of Ethics, Fairness, Accountability, Transparency, Privacy and Employment*.

36. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1-45.

37. Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*(1), 33-36.

38. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.

39. Shneiderman, B. (2020). Design lessons from AI's two grand goals: Human emulation and useful applications. *IEEE Transactions on Technology and Society*, *1*(2), 73-82.

40. Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

41. NRC Web. 2022. *Safety Culture*. [online] Available at: <https://www.nrc.gov/about-nrc/safety-culture.html> [Accessed 9 August 2022].

42. Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

43. Seeger, Anna-Maria; Pfeiffer, Jella; and Heinzl, Armin, "When Do We Need a Human? Anthropomorphic Design and Trustworthiness of Conversational Agents" (2017). *SIGHCI 2017 Proceedings*. 15.

44. Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, Hall P. Beck, The role of trust in automation reliance, International Journal of Human-Computer Studies, Volume 58, Issue 6, 2003, Pages 697-718.

45. Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, *50*(2), 194-210.

46. Taenyun Kim, Hayeon Song, How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair, Telematics and Informatics, Volume 61, 2021, 101595.

47. Pimentel, M. A. F., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, *99*, 215–249. https://doi.org/10.1016/j.sigpro.2013.12.026.

48. McCartney, M., Haeringer, M., & Polifke, W. (2020). Comparison of Machine Learning Algorithms in the Interpolation and Extrapolation of Flame Describing Functions. *Journal of Engineering for Gas Turbines and Power*, *142*(6). https://doi.org/10.1115/1.4045516.

49. Kozyrkov, C. (2021, October 3). *Explainable ai won't deliver. here's why.* Medium. Retrieved August 29, 2022, from https://medium.com/hackernoon/explainable-ai-wont-deliver-here-s-why-6738f54216be