INL/RPT-23-74159
Revision 0

Light Water Reactor Sustainability Program

Explainable Artificial Intelligence Technology for Predictive Maintenance

August 2023

U.S. Department of Energy

Office of Nuclear Energy

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

INL/RPT-23-74159
Revision 0

Explainable Artificial Intelligence Technology for Predictive Maintenance

Cody M. Walker

Vivek Agarwal

Nancy J. Lybeck

Linyu Lin

Anna C. Hall

Rachael A. Hill

Sabid Bin Habib

Ronald L. Boring

Torrey J. Mortenson

August 2023

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.lwrs.gov

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy
Light Water Reactor Sustainability Program

Page intentionally left blank

EXECUTIVE SUMMARY

The domestic nuclear power plant fleet has relied on labor-intensive and time-consuming preventive maintenance programs, thus driving up operation and maintenance costs to achieve high-capacity factors. Artificial intelligence (AI) and machine learning (ML) can help simplify complex problems such as diagnosing equipment degradation to enable more effective decision-making. Benefits will be felt not only within existing analog and digital instrumentation and control, but also work processes, the integration of people with technologies, and most importantly, the business case. Together, these hold the promise to make nuclear power more efficient and reduce costs associated with operations and maintenance. While the AI and ML technologies hold significant promise for the nuclear industry, there are challenges or barriers to their adoption.

Light Water Reactor Sustainability researchers at Idaho National Laboratory—in collaboration with Public Service Enterprise Group (PSEG), Nuclear, LLC—completed development and demonstration of three aspects of AI technologies: performance, explainability, and trustworthiness (represented visually in Figure A).

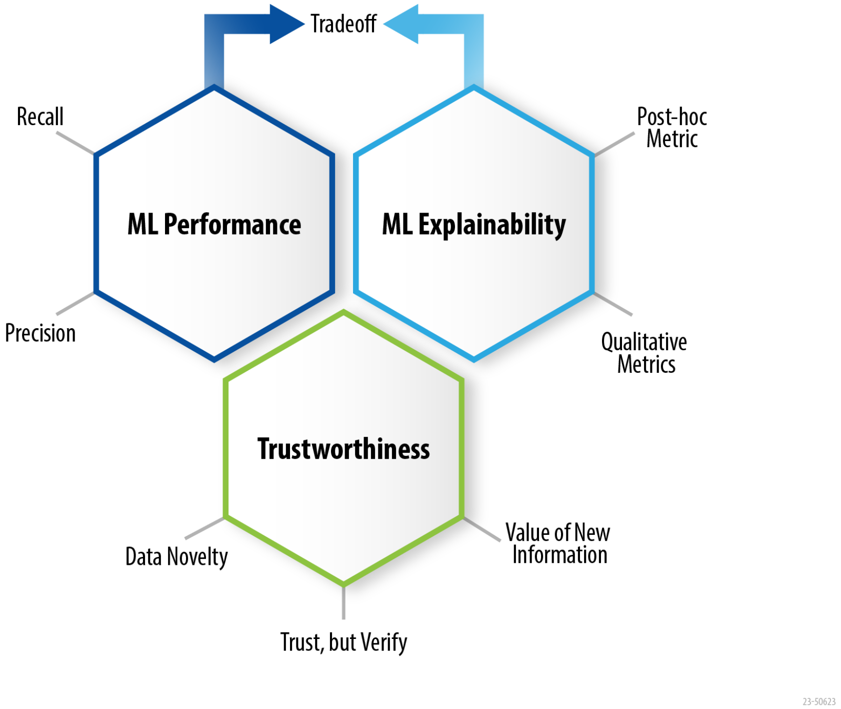


Figure A. Aspects of AI technologies essential for decision-making.

The notable contributions captured in the report are:

- Identify barriers to overcome (categorized as historical, technical, economic, stakeholder, regulatory, and user acceptance) in adopting these new technologies for the industry to realize the full benefits of AI/ML capabilities for long-term economic sustainability.
- Present and discuss the inherent trade-off between ML performance (in terms of accuracy) and explainability, where highly accurate ML methods (such as deep-learning) are the least explainable, and the most explainable methods (such as decision trees) are the least accurate. The trade-off between performance and explainability takes into consideration techniques to develop training

datasets and also concerns around data imbalance. In addition, explainability of AI techniques in terms of transparency and post-hoc metrics are discussed.

- Demonstrate a user-centric visualization that was developed by taking into consideration inputs from PSEG, Nuclear LLC, human factors engineering guidelines, and data scientists. This approach aligns with a human-in-the-loop approach to gain user confidence. The user-centric visualization presents different levels of information and can be tailored as per user credentials. One of the salient features of the user-centric visualization is it presents ML methods with explainability metrics. Battelle Energy Alliance, LLC, holds the user-centric visualization copyright.
- Discuss the trust, but verify framework—a potential approach to building user trust in AI. The framework discusses trust from the human level to the AI level. The fundamental premise of the trust but verify framework is derived from an observation of nuclear safety culture (i.e., nuclear power plant personnel do not rely on a singular source of data to make a decision). This also ties back to the user-centric visualization that presents different levels of information to achieve both explainability and trustworthiness of AI.
- Discuss the findings from the Nuclear Plant Instrumentation Control & Human-Machine Interface Technologies 2023 conference survey that sought feedback from a broader audience about the user-centric visualization app's usability and ML trustworthiness. From the responses it was observed that if the application supplied sufficient information to make the user trust its recommendation, then the participants would likely be comfortable making decisions based on that, even without understanding the underlying details of the algorithm.
- Outline the importance of data novelty and the value of new information in evaluating both explainability and trustworthiness. Novelty detection helps to establish consistency or inconsistency of the new data with respect to the training data. On the other hand, the value of new information could be a part of the user-centric visualization recommendation system that requests additional information be collected to update ML outcomes.

The accomplishments achieved under this research stem from developing innovative solutions that signify advancements in (1) application of AI/ML in nuclear power plants for predictive maintenance, (2) user-centric visualization interface, and (3) quantitative and qualitative measures to achieve explainability and trustworthiness of AI/ML technologies.

The report concludes that the development, implementation, and sustainment of advanced AI-guided technologies will require many different types of expertise for maintenance and regulation. Deployment of AI demands bringing together a truly multidisciplinary team of experts to clearly understand broader societal implications.

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 Jason Tokey of DOE, and Bruce P. Hallbert and Craig A. Primer at Idaho National Laboratory (INL) for championing this effort. We thank Katie Stokes at INL for the technical editing of this report. We thank 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 Public Service Enterprise Group, Nuclear, LLC Monitoring and Diagnostic Center for their valuable technical contributions.

CONTENTS

FIGURES

TABLES

ACRONYMS

AI	artificial intelligence
ARIMA	Autoregressive Integrated Moving Average
CWP	circulating water pump
CWS	circulating water system
DT	delta temperature
EVSI	expected value of sampling information
FNN	Feedforward Neural Network
GMM	Gaussian mixture model
HFE	Human Factors Engineering
HSSL	Human Systems Simulation Laboratory
I&C	instrumentation and control
INL	Idaho National Laboratory
KNN	k-nearest neighbors
LIME	Local Interpretable Model-agnostic Explanations
LWRS	Light Water Reactor Sustainability
M&D	maintenance and diagnostics
MIB	Motor inboard bearing
ML	machine learning
MOB	Motor outboard bearing
NN	neural network
NPIC&HMIT	Nuclear Plant Instrumentation Control & Human-Machine Interface Technologies
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
SHAP	Shapley Additive Explanations
SVR	Support Vector Regression
TERMS	Technology-Enabled Risk-informed Maintenance Strategy
TPE	Tree of Parzen Estimators
UX	User Experience
VoI	Value of Information
WBF	waterbox fouling
WoY	Week of the Year
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNOLOGY FOR PREDICTIVE MAINTENANCE

1 INTRODUCTION AND CONTRIBUTIONS

Over the years, the domestic nuclear power plant (NPP) fleet has relied on labor-intensive and time-consuming preventive maintenance (PM) programs, thus driving up operation and maintenance (O&M) costs to achieve high-capacity factors [1]. As part of the PM strategy, plant systems, structures, and components undergo manual, labor-intensive periodic maintenance checks such as inspection, testing, calibration, replacement, and refurbishment, irrespective of their condition. Predictive maintenance (PdM) strategies, on the other hand, recommend that actions be taken *as required* by the health condition of the systems, structures, and components. To achieve a PdM strategy, condition-based monitoring techniques need to be adopted.

A well-constructed, risk-informed PdM approach (see Figure 1) [2] will take advantage of advancements in sensors, data analytics, machine learning (ML), artificial intelligence (AI), physics-informed modeling, and user-centric visualization approaches. PdM strategies utilize plant assets' current and historical data to develop diagnostic and prognostic models. Diagnostic models identify the current health status of the plant assets. If the diagnosis indicates a potential incipient fault, the prognostic model predicts the time to failure or the remaining useful life, enabling plant personnel to develop a maintenance plan accordingly. These days, NPPs in the U.S. are focusing on transitioning from PM to PdM strategies, one of the cost-effective work reduction opportunities for integrated operation for nuclear [3], in order to achieve long-term economic sustainability in today's competitive energy market [4]. By taking advantage of advancements in sensing, communications, big data analytics, and ML techniques, domestic NPPs are automating and optimizing a number of maintenance activities [5], [6], [7] as part of a larger effort of plant modernization [8].

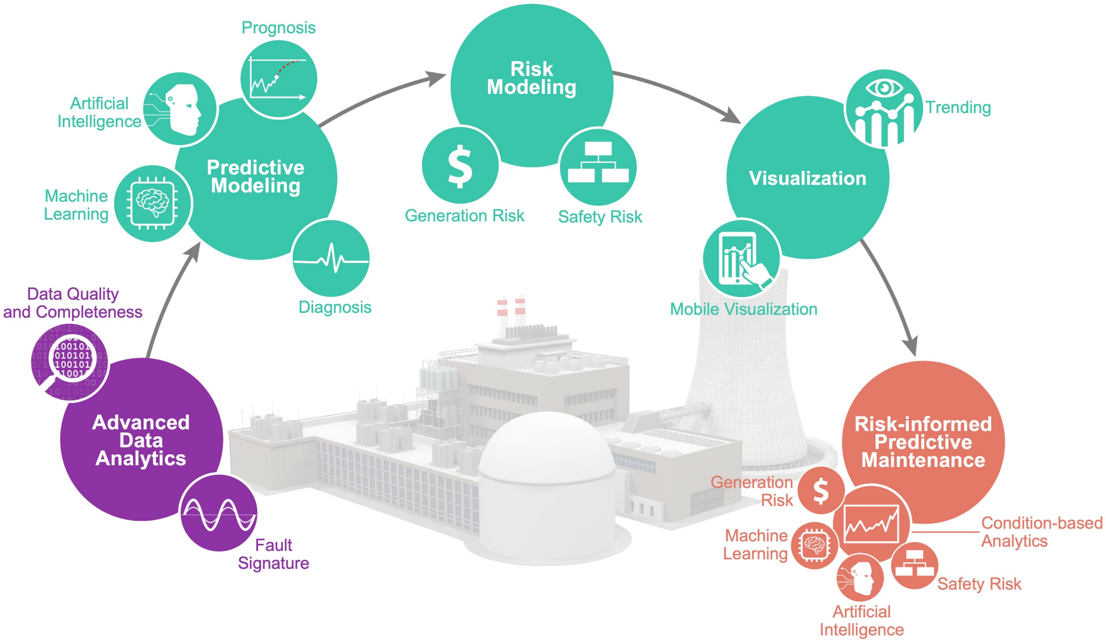


Figure 1. Research and development for achieving a risk-informed PdM strategy [2].

AI/ML is one of the technologies that can help simplify complex problems to enable more effective decision-making. AI/ML is currently being researched in reactor system design and analysis, nuclear safety and risk analysis, and more recently, in plant O&M, especially for advanced reactors [9]. Benefits will be felt not only within existing analog and digital instrumentation and control (I&C), but also work processes, the integration of people with technology, and most importantly, the business case. Together, these hold promise to make nuclear power more efficient and reduce costs associated with O&M. While the AI/ML technologies hold significant promise in the nuclear industry, there are challenges or barriers to their adoption. The challenges or barriers discussed in Section 2 are expected to meet guiding technical requirements as shown in Figure 2. These guiding principles are applicable to the design, development, deployment, and operation lifecycles of AI/ML technologies. For details, see [10].

An ongoing research and development project titled Technology-Enabled Risk-informed Maintenance Strategy (TERMS) under the U.S. Department of Energy's Light Water Reactor Sustainability (LWRS) Program is addressing some of the technical requirements in collaboration with a nuclear utility, including scalability, explainability, and trustworthiness. A federated transfer learning approach has been developed to ensure scalability of AI/ML technologies for risk-informed PdM across plant systems and the nuclear fleet to meet current and future application-specific requirements [11, 12]. However, the developed scalability approach doesn't address the deployment of risk-informed PdM and integration with the plant legacy systems. Explainability and trustworthiness of AI/ML technologies are still open topics of research and development. An initial technical basis addressing explainability and trustworthiness for AI/ML technologies using metrics is presented in [10].

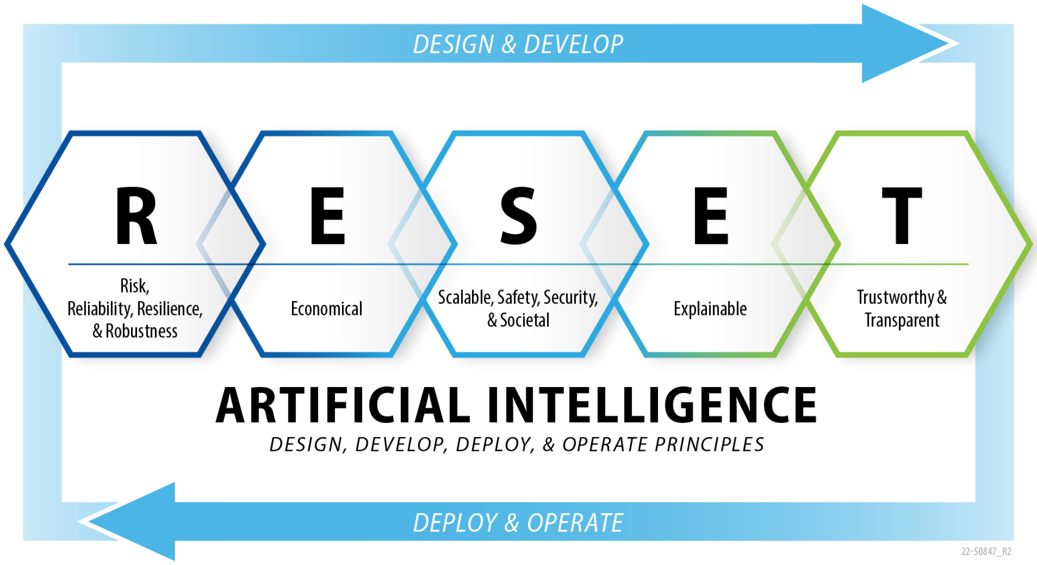

DESIGN & DEVELOP
R
Risk, Reliability, Resilience, & Robustness
E
Economical
S
Scalable, Safety, Security, & Societal
E
Explainable
T
Trustworthy & Transparent
ARTIFICIAL INTELLIGENCE
DESIGN, DEVELOP, DEPLOY, & OPERATE PRINCIPLES
DEPLOY & OPERATE
22-50847_R2

Figure 2. Design, develop, deploy, and operate AI/ML technology requirements.

The primary objective of the research presented in this report specifically focuses on addressing the explainability and trustworthiness of AI/ML technologies to advance the technical readiness and acceptability of these technologies in achieving risk-informed PdM strategy at commercial NPPs. The approach outlined in this report can be adapted to enhance the acceptability of AI/ML in other nuclear applications with a few application-specific modifications. The technical approach 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 technical approach, the circulating water system (CWS) at the PSEG-owned plant sites was selected as the identified plant asset. Specifically, the issue of waterbox fouling (WBF) in the CWS was diagnosed using different types of CWS data.

This report presents discussion on three aspects of AI technologies, i.e., performance, explainability, and trustworthiness, as shown in Figure 3, with specific metrics, a user-centric visualization interface, and human-in-the-loop evaluation to build user-confidence. The notable contributions (below) captured in the report, present detailed discussion on each aspects in the following sections of the report.

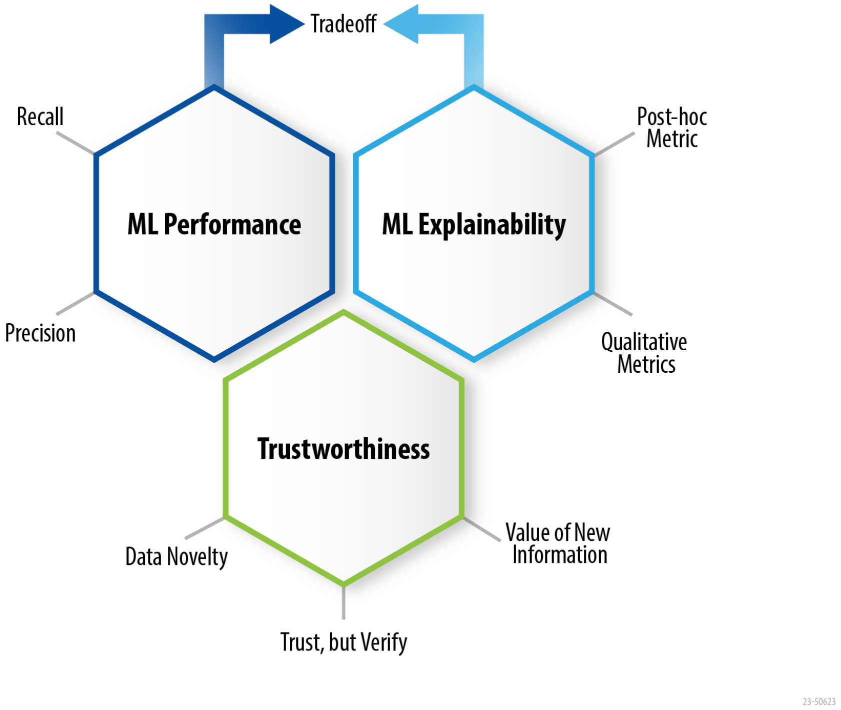


Figure 3. Aspects of AI technologies essential for decision-making.

- Identify barriers to overcome (categorized as historical, technical, economic, stakeholder, regulatory, and user acceptance) in adopting these new technologies for the industry to realize the full benefits of AI/ML capabilities for long-term economic sustainability.

- Present and discuss the inherent trade-off between ML performance (in terms of accuracy) and explainability, where highly accurate ML methods (such as deep-learning) are the least explainable, and the most explainable methods (such as decision trees) are the least accurate. The trade-off between performance and explainability takes into consideration techniques to develop training datasets and also concerns around data imbalance. In addition, explainability of AI techniques in terms of transparency and post-hoc metrics are discussed.

- Demonstrate a user-centric visualization that was developed by taking into consideration inputs from PSEG, Nuclear LLC, human factors engineering guidelines, and data scientists. This approach aligns with a human-in-the-loop approach to gain user confidence. The user-centric visualization presents different levels of information and can be tailored as per user credentials. One of the salient features of the user-centric visualization is it presents ML methods with explainability metrics. Battelle Energy Alliance, LLC, holds the user-centric visualization copyright.

- Discuss the trust, but verify framework—a potential approach to building user trust in AI. The framework discusses trust from the human level to the AI level. The fundamental premise of the trust but verify framework is derived from an observation of nuclear safety culture (i.e., NPP personnel do not rely on a singular source of data to make a decision). This also ties back to the user-centric visualization that presents different levels of information to achieve both explainability and trustworthiness of AI.
- Discuss the findings from the Nuclear Plant Instrumentation Control & Human-Machine Interface Technologies 2023 conference survey that sought feedback from a broader audience about the user-centric visualization app's usability and ML trustworthiness. From the responses it was observed that if the application supplied sufficient information to make the user trust its recommendation, then the participants would likely be comfortable making decisions based on that, even without understanding the underlying details of the algorithm.
- Outline the importance of data novelty and the value of new information in evaluating both explainability and trustworthiness. Novelty detection helps to establish consistency or inconsistency of the new data with respect to the training data. On the other hand, the value of new information could be a part of the user-centric visualization recommendation system that requests additional information be collected to update ML outcomes.

The accomplishments achieved under this research stem from developing innovative solutions that signify advancements in (1) application of AI/ML in NPPs for PdM, (2) user-centric visualization interface, and (3) quantitative and qualitative measures to achieve explainability and trustworthiness of AI/ML technologies.

The rest of the report is organized as follows. Section 2 outlines and discusses different barriers associated with the adoption of AI/ML technologies that need to be addressed to harness full benefits of these technologies in the nuclear industry. Section 3 describes the CWS and its heterogeneous data with different fault modes of interest for PSEG-owned NPPs. Section 4 utilizes the CWS data to demonstrate the trade-off between performance (i.e., accuracy) and explainability of ML methods for WBF condition. Different data sampling strategies along with unbalanced data are used to evaluate the performance versus explainability trade-off. Section 5 presents the development of a user-centric visualization application that presents different levels of information, including diagnosis and prognosis outcomes of ML methods, historical distribution of data, explainability metrics, and trending of salient measurement parameters. Section 6 discusses the trust-but-verify approach that is derived from a nuclear safety culture. Finally, conclusions are drawn and a path forward is presented in Section 7.

2 BARRIERS TO THE ADOPTION OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING TECHNOLOGIES

Despite the industry's excellent track record of providing reliable, safe, and clean baseload electricity for decades, the current business model is unsustainable in today's fiercely competitive energy market [13]. This is because the industry still relies heavily on a large workforce compared to other energy producing utilities that have applied advanced technologies. Put simply, the survival of NPP operations and maintenance, which have remained largely unchanged since the 1970s and 1980s when the plants were commissioned, must keep pace with other energy sectors and, too, harness the dramatic increased efficiencies possible with AI/ML applications.

However, there exist barriers to adopting these new technologies that must be overcome, especially if the industry is to realize the full promise of AI/ML capabilities that will result in significantly lower production costs. The end-state vision of AI/ML adoption is not just nuclear power sustainability but for the industry to flourish. This goal is imperative given the U.S. administration's pledge to decarbonize energy by 2035 [14]. Thus, identifying, and finding ways to overcome barriers to AI/ML adoption is a critical consideration for the industry. This chapter provides a review of these barriers.

We begin with historical barriers that stem from U.S. nuclear power's evolution within the energy landscape, and major events that changed public perceptions and led to its intense safety culture. We highlight technical barriers to AI/ML adoption such as the availability of the granular-level data needed to perform AI, data privacy concerns, and the current lack of AI expert knowledge available at the plants. Next, we discuss potential business case barriers, not least of all, the cost to modernize in this way, but also from the perspective of business stakeholders. Strict regulatory enforcements are in place for nuclear power operations, and we review the 5-year strategic plan for AI readiness recently published by the U.S. Nuclear Regulatory Commission (NRC). Cybersecurity concerns of AI/ML adoption are highlighted. Last, and most importantly, we present barriers to AI/ML adoption in nuclear energy at the user level. As a priority, we lay the importance of User Experience (UX) that includes interface design, explainability, and trust in AI. We describe human nature's resistance to change, and the political and ethical expertise that must accompany AI/ML adoption.

Throughout, the importance of Human Factors Engineering (HFE) in the development and implementation of AI/ML solutions is highlighted. While we categorize the barriers under different headings, we emphasize the overlap and demonstrate their interconnectedness. Figure 4 provides an overview visual of the ways in which the discussion points in this chapter are reciprocally linked. Thus, overcoming barriers to AI/ML adoption in nuclear power will require a high degree of integration and coupling of expertise across multiple domains. No one barrier to adoption should be tackled in isolation, and instead a holistic approach should be applied.

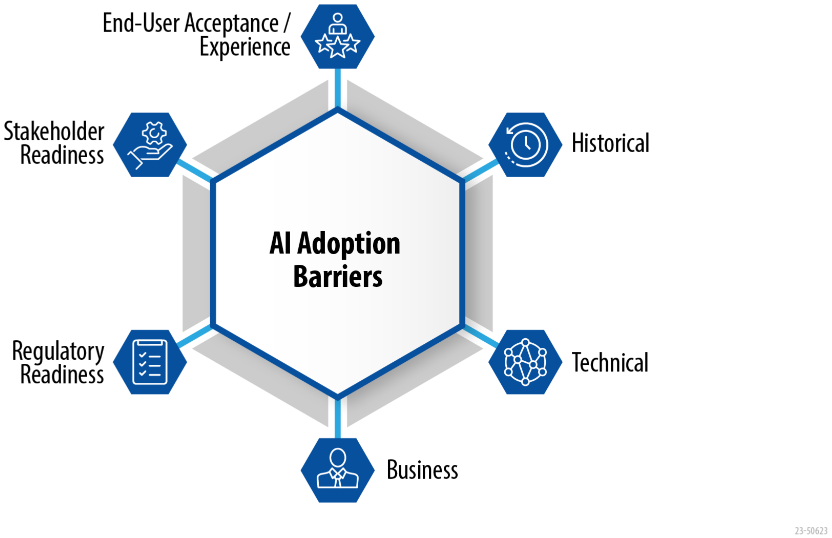


Figure 4. Reciprocal connectedness of barriers to AI/ML adoption in nuclear power

2.1 Historical Barriers

Historically, NPPs have been slower to adopt digital technologies than other energy utilities, and certainly slower than non-energy industries such as manufacturing and automotive which have routinely employed

AI/ML applications for several decades [15]. The reasons for this slow uptake are manifold, but begin with U.S. nuclear power's position within the energy landscape, its safety culture, and its acceptance by society.

U.S. NPPs are among the oldest in the world, with a mean age of 41.6 years. Many of the plants were commissioned in the 1970s and 1980s, and much of the vintage technology of that era is still used today. The nuclear power fleet was not originally built to support upgrades. Further, the events of Three Mile Island in 1979 struck fear into the public consciousness which ushered in a culture of extreme caution and intensive federal oversight [16]. This, followed by the Chernobyl disaster in 1986, effectively placed a moratorium on nuclear development for several decades, and the political factors surrounding its advancement and place in the electricity generation industry have been controversial. Taken together, in comparison to other energy sectors such as oil, gas and, renewables, the nuclear power industry has historically undergone little technological innovation [13], and in instances where it has, the approach has been conservative and delayed.

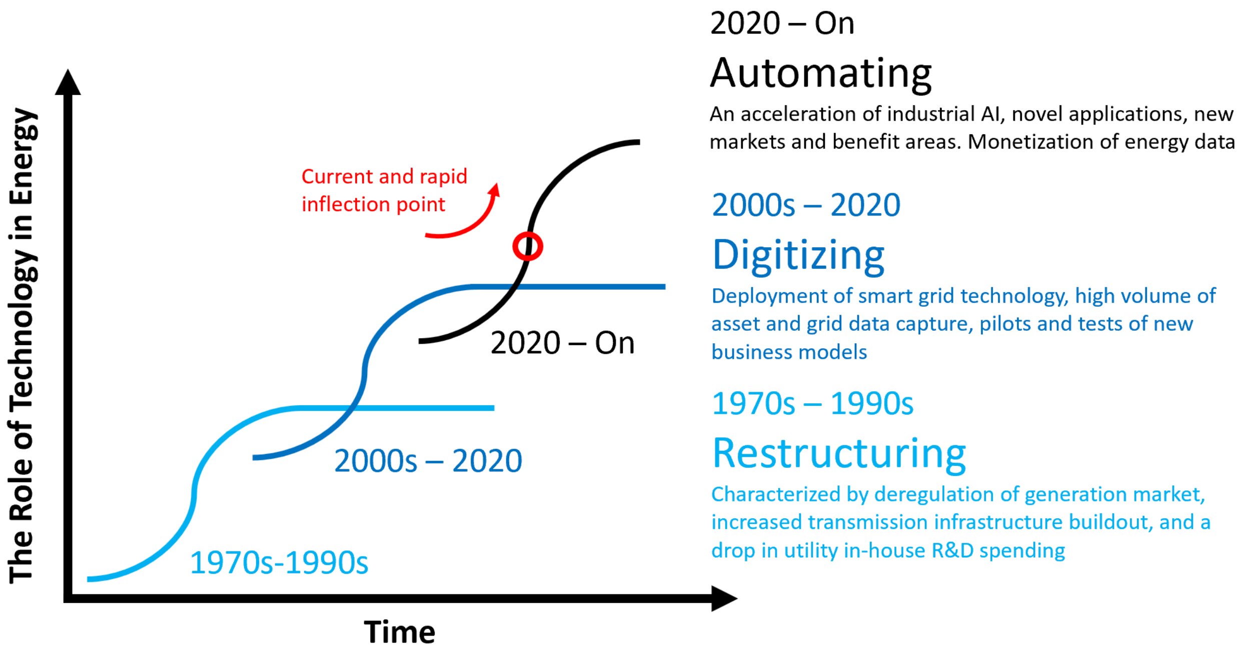

2020 – On
Automating
An acceleration of industrial AI, novel applications, new markets and benefit areas. Monetization of energy data
2000s – 2020
Digitizing
Deployment of smart grid technology, high volume of asset and grid data capture, pilots and tests of new business models
1970s – 1990s
Restructuring
Characterized by deregulation of generation market, increased transmission infrastructure buildout, and a drop in utility in-house R&D spending
Current and rapid inflection point
2020 – On
2000s – 2020
1970s-1990s
The Role of Technology in Energy
Time

Figure 5. Three technological epochs in the power sector (reproduced based on a figure in [17]).

Figure 5 summarizes the technological evolution of the power sector since 1970, highlighting the transformative role that AI/ML applications will play moving forward. The 1970s–1990s saw a restructuring of markets and an infrastructure buildout that brought about a drop in research and development. This era tracked with the introduction of renewable energy sources. The 2000s–2020s ushered in the epoch of digital transition, including investment in smart technology and deployment of data capture devices. Most recently, since 2020 we began the automation era boosted by data engineering and AI/ML technologies. The automating epoch contains new clean energy markets and new business models.

In recent decades nuclear power utilities have been engaging with I&C digitalization and modernization of the existing fleet to varying degrees. Each upgrade poses significant challenges to the plant including financial investment and training considerations. Further, most technological advancements attract a regulatory audit and, in some cases, a license amendment request. Thus, cost concerns, lack of expertise, and regulatory issues surrounding new technology adoption are not new, and have historically been the industry's chief barriers to modernization.

Recent findings by Hall and Joe [18] suggest that industry's attitudes toward some of these perceived barriers may be declining while others are increasing. The authors asked personnel from utilities and the nuclear industry, including vendors and researchers about perspectives of control room modernization. They surveyed individuals in 2012 and again in 2022 to gain insights into whether attitudes had changed over the course of a decade. Cost remained a primary concern and increased from 2012. However, despite being the foremost barrier 10 years ago, concerns about the regulatory approval process declined by a whopping 30% and was among the lowest ranked barriers in 2022 (Figure 6). With respect to AI/ML industry adoption, these findings are promising. Thus, despite historical reservations about NRC-approval there currently exists a shared sentiment across utilities and research agencies alike that regulatory approval should not pose an insurmountable barrier to AI/ML adoption in the plants moving forward.

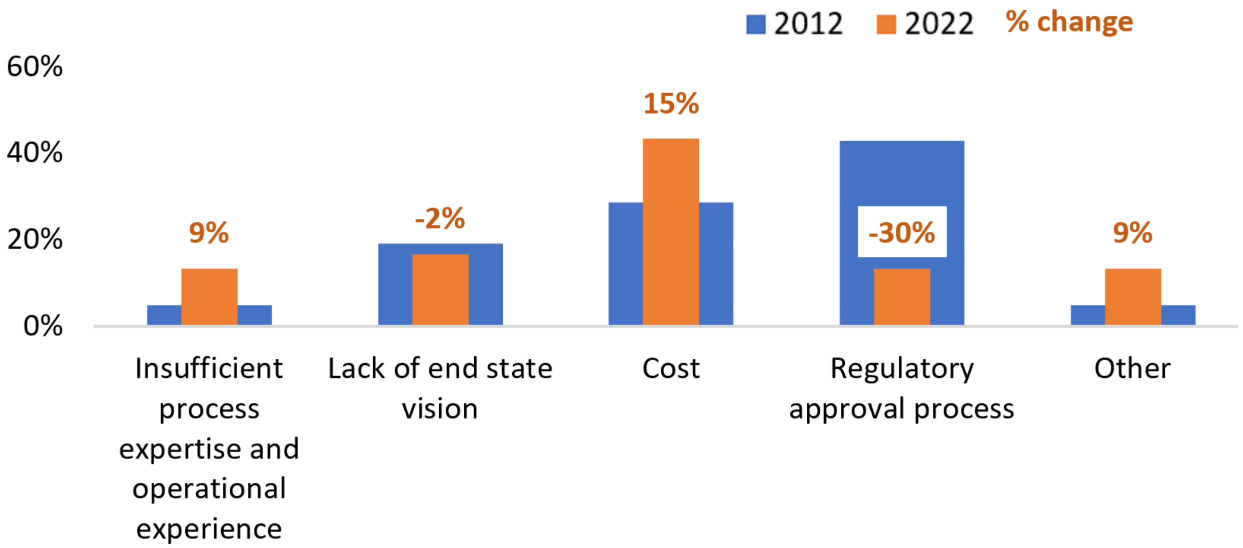


Figure 6. Foremost barriers to control room modernization (Reproduced with permission from [18])

Taken together, the historical underpinnings of analog equipment, stagnation in development for several decades, dominant regulatory oversight and events such as Three Mile Island and Chernobyl have led to the industry being the safest energy generation source and with a strong safety culture. This is a crowning achievement and one of which the industry can be proud. However, this has also led to slow installation of technological advancement over the decades, including AI.

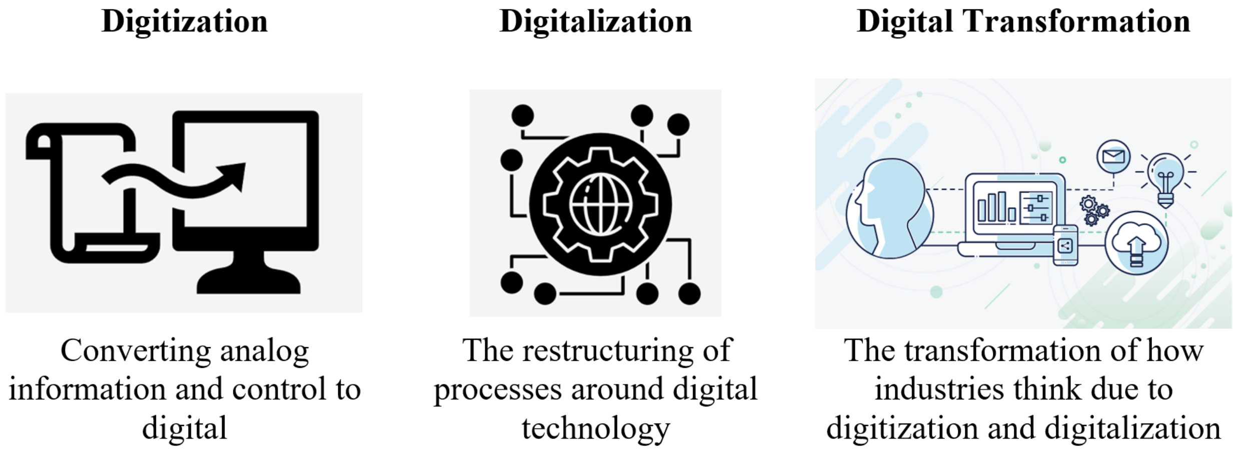
2.2 Technical Barriers

There's a saying within the AI/ML community, “Garbage in, Garbage out” [19]. Insofar as AI's primary functions are for the system to learn knowledge from data, then correctly interpret that knowledge and act accordingly, the granular-level data necessary to support knowledge learning is a technical barrier that must be overcome. Here, we parse out issues pertaining to data accessibility, data quantity, and data quality as a prerequisite for successful AI/ML application in nuclear power.

2.2.1 Data Accessibility

Often, the real-time plant data necessary to perform AI/ML are rarely available. Indeed, many AI/ML applications are developed with simulated data from high-fidelity models [9]. Further, should these data be collected in the first place, gaining access is not a trivial matter because they exist in a closed loop system within the plant and are closely guarded. This is a function of significant data privacy and cybersecurity concerns that NPPs must contend with, and written agreements for AI/ML researchers to access the data can take several months. Further, algorithm optimization requires continuous availability of real-time data from plant process systems and other installed sensors. Ideally, this occurs in an uninterrupted manner. Brokering contracts that provide AI/ML developers with perennial NPP data access will have to be made easier to realize the technology's full potential.

Within the evolution of digitized information, the data must be stored in an accessible format. For example, information stored in portrable document formats is not readily obtainable for use by AI/ML analytics compared to other digitized formats. Digitization refers to the conversion of analog information into digital form. An example of this in nuclear power operations is the transformation of paper-based procedures into computerized procedures conducted on a tablet of some kind. With this new form of technological infrastructure in place, plants can engage in digitalization, which uses digitized information to restructure and improve business processes [20]. Digitalization involves optimizing interactions between systems, between people, and between systems and people using digital communications. This new format yields insights and possibilities that would have otherwise been impossible with analog systems. Together these innovations enable digital transformation; a fundamental change in thinking and the creation of new, more efficient ways to perform daily operations [21]. An example of this process of digital transformation is illustrated in Figure 7.



Digitization

Converting analog information and control to digital

Digitalization

The restructuring of processes around digital technology

Digital Transformation

The transformation of how industries think due to digitization and digitalization

Figure 7. Industry pathway to digital transformation (image courtesy of [21]).

Finally, many plant tasks such as work orders, corrective action reports and others could be optimized by easy access to textual information stored in text-based documents. However, the vast majority of available textual information is currently stored as unstructured text, such that keyword searches still require the user to read the file once located. This process is labor intensive, error-prone, and costly. AI/ML applications such as Natural Language Processing can locate relevant textual information quickly from vast volumes of text-based documents and find patterns that are not readily apparent to users who perform a visual search. The transformation of free text into normalized, structured data can then be fed directly into ML algorithms that support decision-making during plant operations and maintenance.

2.2.2 Data Quantity

The quantity of data required is a serious consideration. In recent years, many forward-thinking NPPs have made large investments into digitizing infrastructure by deploying smart sensors and data tracking devices for information that was either previously manually recorded, or not recorded at all [9]. For example, the risk-informed predictive maintenance application detailed in this report, depending on its application, will rely on large volumes of data such as motor vibration and temperature, system condenser, and turbine and vacuum values [10]. Other recent digital upgrades to I&C have included smart gauges, electronic work packages, and dynamic instructions [22]. This has produced new digital data from disparate origins that, when connected, have made possible the creation of new sources of knowledge. Together, these devices create a digital environment from data capture that can yield new and valuable insights using AI/ML. Recent advancements in sensor hardware, cloud-computing, and bandwidth have lowered costs to put these data capture technologies in place [23]. However, while some utilities have made great strides updating their data information infrastructure, there is currently still not enough digitized and relevant plant data available to feed and validate most high-fidelity AI/ML applications [9]. Further, NPPs will have to invest in prolific data capture resources, software engineering, and sophisticated software architecture to meet this demand.

2.2.3 Data Quality

Even with ready access to high quantities of relevant plant data, AI/ML's ability to function correctly is also dependent on the quality of that data. Thus, to arrive at quality decisions, there must be clean, accurate, relevant, reliable, and contextualized data. Data must be recorded faithfully by devices, and errors corrected before further data processing occurs. Incorrect, mislabeled, and duplicate data values must be removed, and incomplete or corrupted data must be handled to produce valid information. Further, the cleaning process must be performed in a consistent manner to ensure data reliability. There must be checks and balances in place to verify that data cleaning was performed properly, with course-correct mechanisms in place via feedback loops. Last, the data may need to be transformed into a uniform and contextualized format, especially across multifield, multidimensional heterogenous datasets from disparate sources [9]. Taken together, the data engineering and safe data storage that must occur before AI analytics and model training is not a trivial issue and will require substantial backend investments by the utilities to ensure suitability. Figure 8 details an example of a big data pipeline process and the developers involved at each stage. A full discussion of the data processing required for model development pertaining to risk-informed PdM can be found in [24].

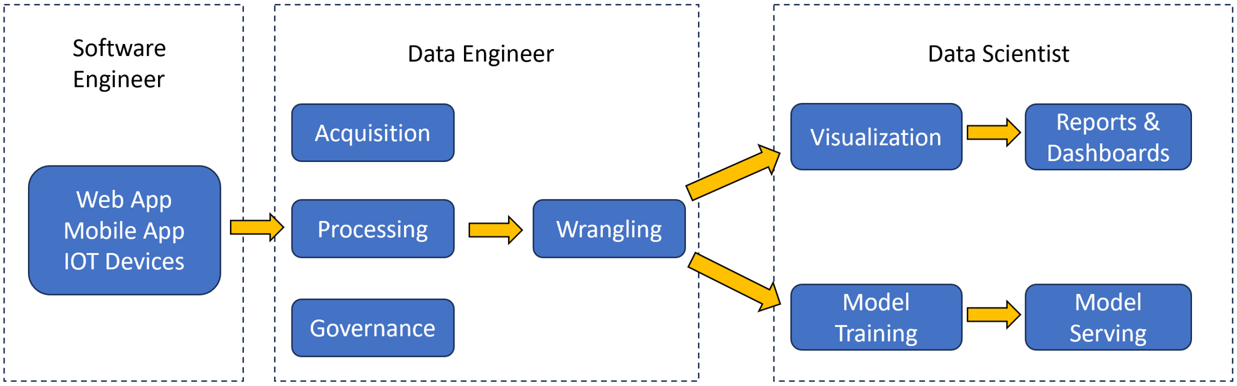


Figure 8. Data engineering and developer tools for big data (from [25]).

2.2.4 Level of Required Technical Expertise

The last technical barrier described here relates to the level of required technical expertise, which can be broadly defined as the level of AI/ML technical maturity that personnel possess within a utility. In parsing out technical maturity, skills expertise is implicated in two ways. The first is the existence of expert AI/ML knowledge and competence. The second is the acquisition of new skills required with AI-augmented and newly created AI positions.

The existence of expert AI/ML knowledge and competence becomes a barrier to adoption when the plant lacks employees with the specialized technical background to comprehend the AI/ML model analytics. This issue is particularly salient in the nuclear industry because O&M decisions often carry significant safety and financial consequences, as well as legal consequences. NPP operators hold licenses and bear legal responsibility for plant operations, whether their decisions are AI-guided or not, and so an understanding of how the algorithm arrives at its conclusions is paramount.

Explainable Artificial Intelligence (XAI) has become an important feature of AI/ML technologies and is an essential component in earned trust by the human user [26]. XAI refers to the system's ability to communicate or explain how it arrived at its decision or algorithmic output [27]. Indeed, there exist consumer laws dictating that algorithmic output must be explainable to users. For example, the Equal Credit Opportunity Act, passed by the U.S. Congress in 1974 states that any adverse actions against those seeking credit must be explained, including actions informed by algorithmic decision-making [28]. However, for NPP personnel to sufficiently understand a technical explanation based on complex computations performed upon plant data, this requires a degree of model expertise and extensive background knowledge [29]. Should there be a shortage of AI/ML talent in the plant, model training and recruitment will be required in every department that employs these technologies. A discussion of XAI as it pertains to the nuclear industry is documented in [10].

Last, there is no question that the introduction of AI/ML will change the nature of current job functions and create new, different jobs altogether that will require augmented and new employee skillsets. Indeed, in a recent report from the World Economic Forum, it is estimated that across the globe, 44% of workers' skills will be disrupted in the years leading up to 2027, in part because of AI applications, and 60% will require additional training [30]. Analytics, creative thinking, and understanding of how to use AI and big data were the top-ranking skills training in terms of importance. This is apparent as AI/ML technologies assist with automation and decision-making in industries ranging from finances, automotive, and manufacturing [15]. Human-machine collaborations will become commonplace for most job descriptions including those in NPPs, and while this will allow job functions to become more streamlined, the original mandate of the employee will be displaced by something else. With the adoption of AI/ML technologies, the nuclear industry will have to be cognizant of creating an environment that ensures the correct skillsets are being developed in personnel.

2.2.5 Governance

AI/ML model training is stochastic in nature which makes verification and validation tasks challenging. Also, systems using AI/ML models are difficult to audit and certify because of their black-box nature. These concerns are further challenged by intrinsic biases in AI models such as reproducibility bias, selection bias (e.g., races, genders, color), and reporting bias (i.e., results that do not reflect the reality). They also appear to be vulnerable to cybersecurity threats as AI systems can misbehave when untrusted data are given, making them insecure and unsafe.

While the above mentioned concerns are true for many AI systems, the application of AI in nuclear has to deal with additional concerns of (1) integration of AI system with NPP legacy systems without disturbing the intended functionality of the system; (2) expanding current infrastructure at a NPP in terms of sensors, communication, data repository, computation resources, cyber security, and human resources to effectively accommodate the new AI systems; and (3) ensuring regulatory compliance.

While this is not a comprehensive list of concerns related to potential AI/ML applications in nuclear, they warrant developing and implementing a governance framework. The structure and functionality of the established framework would be to provide an oversight ensuring concerns associated with the AI/ML system throughout their lifecycle (i.e., design, develop, deployment, and operation) are monitored and addressed in a timely manner. The AI governance framework is expected to be based on guiding principles, to be defined by a collaborative effort of AI/ML developers, cybersecurity experts, information technologists, stakeholders, regulators, and whoever else required. There are several guiding principles related to AI governance in literature for consideration. For details, see [31], [32], [33].

2.2.6 Cybersecurity

The success of AI/ML applications in NPPs is predicated on access to readily available, highly interconnected electronic data streams from disparate and varied sources. Thus, while AI/ML holds promise to significantly increase efficiency and lower costs, the shift to digitalized information leaves plants more vulnerable to cybersecurity breaches. This constitutes a serious safety concern and, by extension, a regulatory concern. Cybersecurity within NPPs is at the forefront of discussions surrounding digitalization and AI/ML, and should be considered during the formative, planning stages of the technology. Smart devices and intelligent agents generate efficiencies but also increased risk [34]. Keeping digitized nuclear power assets safe from cyberattacks is a matter of national security [35]. After a spate of cyber disturbances in the 2000s, the NRC mandated an ordinance that requires NPPs have a cybersecurity plan in place that meets the Commission's approval (Title 10 of the Code of Federal Regulations, as outlined in NRC Regulatory Guide 5.71; US NRC, 2009).

Scientists and engineers at INL are at the forefront of cybersecurity research and development for nuclear power systems that include technological solutions within cyber defense. Importantly, HFE scientists at INL can examine vulnerabilities in human-AI interactions using the Human Systems Simulation Laboratory (HSSL), such that the risk-informed PdM outlined in this report can be developed within the existing framework of a utilities' NRC-required cyber program. An outline of cyber-related research activities conducted at the HSSL can be found in [18].

2.3 Business Case

Cost has always been a significant barrier to modernization, but it has become an increasing concern in the last 10 years [18]. As with other types of technological upgrades, the upfront investment and ongoing data expertise necessary for the success of AI/ML technologies likely remains a chief barrier to adoption. Stakeholders must perceive guaranteed long-term benefit in conjunction with a friendly regulatory environment to offset the initial investment. Further, effectively overcoming this barrier will likely rely on each of the business stakeholders possessing a clear and unified end-state vision of AI/ML throughout the plant. Within the industry, there are mixed feelings about whether the perceived development and implementation costs are worth the benefits [9]. There are concerns about the timeliness of received benefits, and the risk-aversive nature of the nuclear power industry likely adds costs to AI/ML adoption

Utilities perceive that each potential AI/ML application must be considered on a case-by-case basis, and the risk of failure or malfunction and anticipated efficiency improvements over time must be weighed against future value. However, large scale transformation to a digitized infrastructure is precisely the type of seamless digital environment that will support plant scale AI/ML deployment and, by extension, realize the technology's full potential. Finding solutions to navigate and overcome these competing tensions (i.e., case-by-case AI/ML versus scale deployment) is part of the business barrier.

Critical to the business case is that clear methods must be in place to identify key performance indicators that measure return on investment. For example, it is currently challenging to capture relevant and credible cost-savings data from AI/ML implementations in component monitoring, the benefits of which are realized months or even years down the line. Nonetheless, component monitoring for early fault detection and predictive maintenance has been identified as one of the key areas of nuclear power that AI/ML technologies will benefit from the most, allowing for optimization of resource allocation and cost-savings [9].

2.4 Regulatory Readiness

The nuclear power industry faces stricter regulatory enforcement than most business enterprises, and any application of AI/ML technologies will require a cooperative and collaborative approach with the U.S. NRC, the federal oversight agency responsible for ensuring safety requirements are met. To this end, in their document titled 'Artificial Intelligence Strategic Plan: Fiscal Years 2023–2027,' the NRC stated that its vision is to *"continue to keep pace with technological innovations to allow for the safe and secure use of AI in NRC-regulated activities, when appropriate"* (NRC, 2023, p.2-1). In this section, we will also consider cybersecurity concerns.

2.4.1 NRC Readiness

The introduction of new digital technologies to NPPs, including AI/ML to plant processes, might attract regulatory oversight. Although AI/ML is currently being explored in the optimization of non-safety systems, should the upgrade include safety-critical systems, the plant will have to submit an amendment application to the NRC for approval. Without approval, the plant cannot maintain its license to operate. The NRC's 5-year strategic plan for AI is essentially a document outlining the agency's preparations for readiness to review licensee submissions that employ AI technologies. This is not an easy task because regulating a rapidly evolving AI/ML landscape presents novel and unique challenges.

The NRC strategic plan includes five goals:

1. Ensure NRC readiness for regulatory decision-making
2. Establish an organizational framework to review AI applications
3. Strengthen and expand AI partnerships
4. Cultivate an AI-proficient workforce
5. Pursue use cases to build an AI foundation across the NRC.

The first goal is the most important and involves establishing a secure AI/ML decision-making framework that is up-to-date and that stems from a sound technical basis. This will form the basis of inspection procedures and oversight policy. The need for flexibility is recognized. Technical AI topic areas that will be present in the decision-making framework for regulatory approval include bias, security, risk analysis, model maintenance, and data quality. It will also review XAI, trustworthiness, and ethics of the algorithms-—we discuss these important considerations in Section 2.5, 'User Barriers,' below.

Goals 2–5 are in service of this first strategic goal. The second goal relates to the creation of internal NRC agency coherence across departments in reviewing applications. To accomplish their third goal, the NRC is calling for regular and early engagement from applicants considering AI/ML, in conjunction with partnerships from national and international subject matter experts. The agency has been regularly conducting AI workshops that engage utilities, industry, intergovernmental entities as well as the public to discuss data science and AI that is being considered for applications in the nuclear industry. The fourth goal is to build core AI-competencies in NRC personnel and attract AI talent to the agency who are poised to perform effective regulatory reviews. Last, the fifth goal is to develop an end state vision of AI from which NRC-regulated activities can be guided, including refinement of AI policies as feedback from industry use cases and international expertise becomes available. Taken together, these goals will culminate in AI technical readiness for regulatory reviews.

On the left-hand side of Figure 9, taken from the NRC 5-year strategic plan, the pyramid depicts the data structure that must already be in place before a utility embarks on data science and AI technologies. The NRC highlights core data competencies already in place stemming from extensive experience regulating the lower levels of this pyramid. Its strategy with AI will build upon these capabilities. The right-hand side of Figure 9 demonstrates that machine learning techniques (like predictive maintenance), as the NRC views it, is a subset of AI with its foundations in data science.

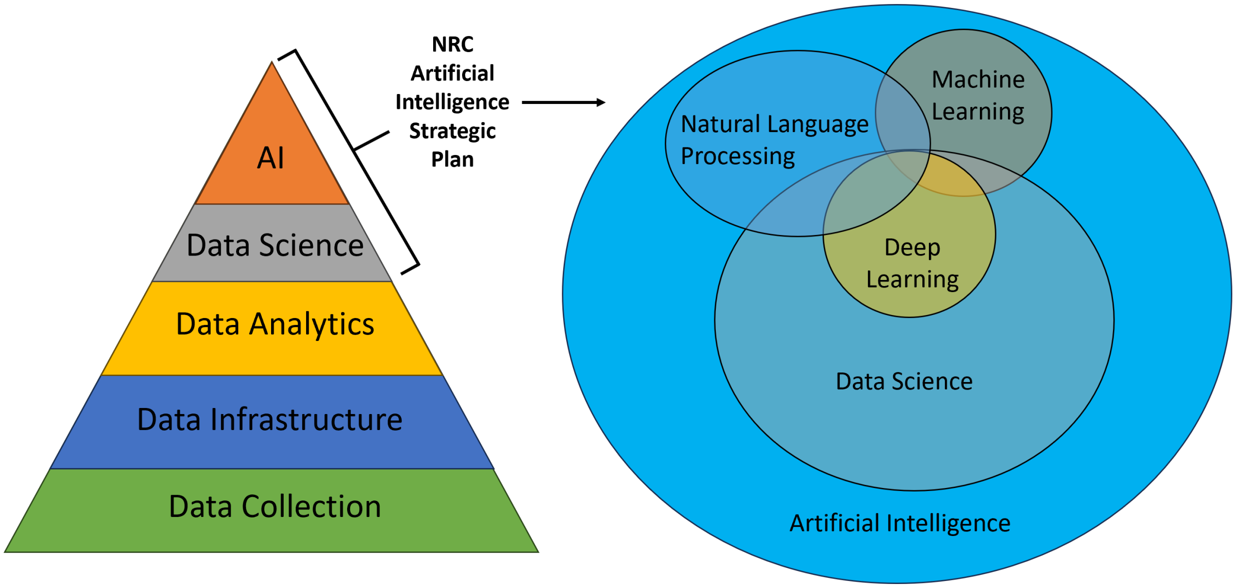


Figure 9. AI hierarchy and relationship with the NRC AI strategic plan.

To this end, the NRC stipulates that notional AI and different degrees of autonomy will require different levels of regulatory scrutiny. This stems from the fact that some AI applications serve automation, while others serve autonomy. The difference lies in the level of responsibility the intelligent agent has in decision-making processes; less human intervention requires greater regulatory scrutiny. These levels are listed below from Level 0 (100% human decision) to Level 4 (100% machine decision). The risk-informed PdM innovation outlined in this report would likely be categorized as Level 2 Collaboration whereby algorithms make recommendations, but these are vetted and the action remains with the human decision-maker. The level of regulatory scrutiny will also depend on whether the AI/ML affects safety or non-safety plant systems.

- Level 0: AI Not Used
- Level 1: Insight (Human decision-making assisted by a machine)
- Level 2: Collaboration (Human decision-making augmented by a machine)

- Level 3: Operation (Machine decision-making supervised by a human)
- Level 4: Fully Autonomous (Machine decision-making with no human intervention).

2.5 User Barriers

UX is another important factor to consider in the successful adoption of AI/ML technology within nuclear power operations and maintenance. Existing plant personnel must be able to observe immediate benefits, and likeability of the technology plays a big role within user approval. HFE scientists must work closely with data engineers and plant engineers to ensure that the human-AI interface is designed with sound HFE principles that result in a product with which the users are apt to engage. Embedded within UX are interface design, explainability, interpretability and trustworthiness of the AI/ML, all of which are requirements for user adoption.

2.5.1 User experience

The human-AI interface is the way that personnel interact with AI/ML technologies in order to carry out their duties. As a basic principle, the interface must use a presentation format that is consistent with the task functions the user is to perform [36]. Other design principles include being able to hold the user's attention, minimize errors, and afford non-experts an understanding of how the system works. The interface undergoes early revision in parallel with the AI/ML system design, and the new, improved design is validated through an iterative process. The U.S. NRC has developed general HFE recommendations for interface review including physical and functional characteristics that should be present in the information display [37]. Safety and usability are the chief priorities.

A rich HFE theoretical literature is devoted to interface design, with specific recommendations made for process control industries such as nuclear power operations [38]. Ecological interface design uses cognitive psychology to inform ergonomic visual representations that best serve humans working with complex sociotechnical systems [39, 40]. Two NRC documents offer guidance on HFE review criteria for interfaces including design features [41] and interfaces to automatic systems [37]. Together, the NRC makes clear that the display must both fulfill the needs of the system and users with safety as a priority.

Important for UX in digital displays is the stylistic elements present because these can impact interface performance as well as user preference [18]. Hossain and Zaman [42] outline fundamental design considerations that together create an inviting environment for the user, and an interface that aids user understanding. When producing a clear visual understanding of the system's behavior, the authors urge that the following factors be taken into consideration:

- Screen layout
- Color
- Graphics and pictures
- Display text
- Value representation
- Alarms and events

- Navigation and controls
- Page hierarchy
- Operational security.

Some of these features are specific to interfaces used for nuclear operations (e.g., alarms and events, operational security). However, several features that are relevant to interface design of the ML-based technology described in this report will be considered briefly. The screen layout should reveal key information in the sequence that human operators naturally scan any screens (i.e., from top left to right, then down). Color should be considered carefully because red typically conveys stop/emergency whereas green conveys start/safe conditions. Color use must also comport with nuclear power conventions. In terms of graphics, according to the High Performance Human-Machine Interface Handbook, a good graphic should have a grey background (and not blue, for example [43]). Data values representing key system parameters should be presented graphically and not as text. Taken together, it is important for AI-interface designers to understand that UX is about human perception as much as data representation.

XAI is another critical component to UX and has been a major topic of discussion surrounding AI/ML since the 1980s. Just as humans in important decision-making roles are asked to explain their decisions, XAI serves to hold machines to the same standard, especially when the underlying computations may be opaque to the user as with highly complex sociotechnical systems such as nuclear power.

Although often used interchangeably in the computer science literature, some scholars of psychology propose that interpretable AI differs from XAI and that they represent two distinct psychological constructs. For example, [29] argues that algorithmic interpretability is whether or not a human can meaningfully understand the accuracy of the model prediction for the task at hand. In other words, unlike XAI which is concerned with explaining the process of decision-making, interpretability pertains to the system's decision-making accuracy in fulfilling its purpose. The author illustrates the difference between explanations and interpretations with an example of the "Check Engine" light on a dashboard: the explanation might be detection of a faulty driving system, but the interpretation for the driver is to take the car to the mechanic.

Users of AI/ML technologies seek both explanations and interpretations simultaneously, and both are central components of a favorable UX. They are requirements for new systems. Individual differences of proficiency for target users dictates how much explanatory information the system should provide in parallel with the algorithmic output. Reference [29] notes that, whereas humans make decisions based on the simplest representation available, or the gist—the essential meaning of the model output (interpretation)—AI/ML models arrive at decisions via methodical programmatic verbatim processes. Thus, highly explainable algorithms must provide detailed descriptions for the rationale behind the model's outcome, whereas highly interpretable algorithms must afford meaningful mental representations that allow the user to understand the model's outcome in context, which, in turn, supports high-level decision making.

Further, there is a negative correlation between algorithmic performance and explainability in that users must weigh the trade-off between model accuracy and an explanation of the model's decision process [44]. As outlined in [10], an important question within nuclear modernization is the extent to which reactor operators must grasp completely the underlying control logic and behaviors of AI/ML technologies. Explainability and interpretability are both difficult to measure [26]. However, given their importance in user adoption, AI/ML developers must work closely with HFE scientists to develop applications that optimize UX. A key aspect of this optimization will be soliciting user feedback in an iterative fashion to improve explainability and interpretability. Human-centered AI is an approach to AI/ML development with the requirement that user needs be built into the ML model's design [45]. Human interaction is emphasized, along with a user-centric design that supports cooperation between human user and artificial agents. Altogether, these design principles increase human control and trust in the system.

2.5.2 Ethical barriers

In addition to XAI, any lack of transparency surrounding how the algorithm made its determinations, or any user misunderstanding can contribute to ethical concerns. This stems from fears that unconscious biases in AI developers made their way into the application. Indeed, while on the one hand AI-guided decision-making is perceived as more objective than that of humans, any subjectivity or bias from the developer may have been embedded into the algorithmic learning [46]. Other ethical risks include users not wanting the responsibility of applying judgment to AI creations and ambiguity surrounding where responsibility lays should accidents occur. This is particularly relevant for the nuclear industry because staff also bear legal responsibility for accidents.

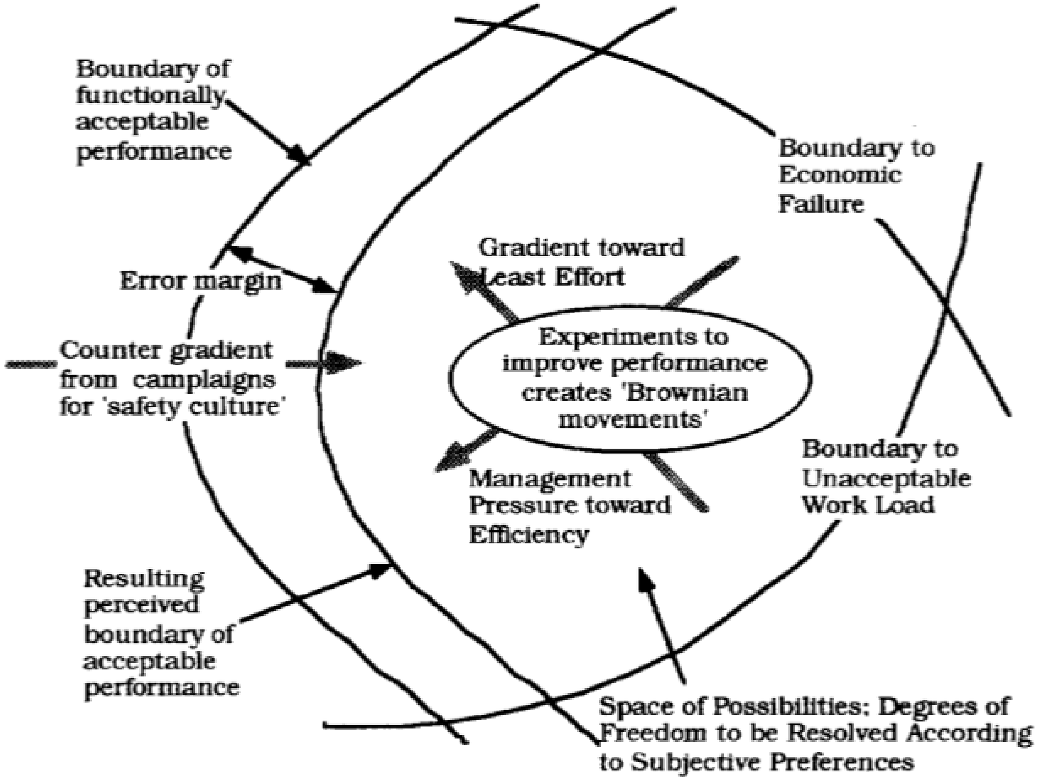
Further industry concerns about privacy and surveillance ethics have been expressed, for example, from AI-guided technology that tracks plant personnel location (i.e., The Global Positioning System) in real time or that uses face recognition for permissions (i.e., cameras). Last, the far-reaching consequences of AI are still not yet known and pose significant ethical and societal concerns for an industry in which errors can have far-reaching implications. Taken together, the development of AI-guided applications must address user ethical concerns early in the process to increase trust in the technology. The implementation method must feature a comprehensive ethical risk mitigation plan. Taking the time to overcome ethical user barriers to adoption is as important as UX.

2.5.3 Organizational readiness

Another barrier to adoption is organizational resistance to change, which is a function of organizational readiness. This is, in part, due to the safety and risk-aversive culture of the nuclear power industry described earlier, which generates a general lack of trust in single sources of information as well as the overall efficacy of predictive models. Organizational readiness is variable across the industry, with some utilities more embracive of technological innovations than others.

New AI/ML applications must be developed with, and validated by, current plant personnel. Regulatory approval is contingent upon a favorable reception from plant operators and those who work closely with existing systems that the new technology seeks to augment or replace. However, this presents a conflict because familiarity with legacy technology or earlier plant processes creates a user bias: the perceived appropriateness of the advanced AI-guided systems is reduced with more experience using the older system [47]. Thus, earning the trust of the target user is paramount for engagement. A review of human-centered AI as a function of willingness to trust the system is provided in [10].

In addition, the introduction of anything new is by definition a political act, because it causes change. And, as with most other organizations, the nuclear power industry experiences resistance to that change. Change causes disruptions at different levels. For a nuclear power operator, who is responsible for maintaining safe and efficient operations at the plant, the introduction of a different work process may initially result in errors and perceived or real punitive action against the employee. In addition, learning new systems takes time, and is cognitively more burdensome than established practice. There exists a tension within human-AI collaboration in that, despite the known efficiency benefits, a culture of organizational fear might surround the new technology in that it will ultimately replace human labor. Therefore, a lack of trust in AI/ML may stem from job loss concerns, and not in the specific application per se. Utilities must strive to create a culture friendly to innovation, with reassurances that adequate time will be given for training and to master the new technology, and that job security is a prerequisite. Moreover, this is compounded by the fact that technological changes outpace managerial changes at an exponential rate, which in turn revolves at a faster pace than those in organizational legislation and regulation [48]. Given their dynamic and necessary interaction to exploit AI/ML to its full potential, common ground must be found to balance these ingredients and ensure proper communication.



Boundary of functionally acceptable performance
Boundary to Economic Failure
Gradient toward Least Effort
Error margin
Counter gradient from campaigns for 'safety culture'
Experiments to improve performance creates 'Brownian movements'
Management Pressure toward Efficiency
Boundary to Unacceptable Work Load
Resulting perceived boundary of acceptable performance
Space of Possibilities: Degrees of Freedom to be Resolved According to Subjective Preferences

Figure 10. Worker behavior in the presence of organizational gradients [48].

Rasmussen [48] outlines different organizational pressures within each department and staff member that interact simultaneously and affect decisions. Figure 10 depicts the presence of strong forces in opposing directions that, together, likely result in staff migration toward the boundary of acceptable performance. Management pressure toward efficiency (e.g., advanced technology adoption) is one force which stems from aggressive market realities—this flows against the gradient toward least effort. The boundary of functionally acceptable performance opposes the boundary of unacceptable workload. The human workers must achieve their work objectives within constraints such as nuclear's safety culture and administrative and functional restrictions. However, there exists latitude for workload, efficiency, risk of failure, an individual's joy of exploration, and adaptability on the part of the worker. The Brownian movement in the center of the figure represents staff behavior trending toward boundaries in all directions depending on their collisions with strong currents from all directions. Put simply, beyond individual motivations, where users of AI-guided technology ultimately lay within workplace boundaries depends on the outcome of the worker 'effort gradient' meeting management's 'cost gradient.'

2.5.4 Individual differences to technology adoption

The technology acceptance model is a framework to examine and understand user perceptions that best determine technology adoption, as well as longer-term usage [49]. The model is based in psychological research and attempts to combine principles from the theory of planned behavior and the theory of reasonable action that together explain users' behavior toward technology usage. Figure 11 shows the model framework, in which perceived usefulness of the technology (i.e., enhance job performance) and perceived ease of use (i.e., free from effort) in combination form a usage attitude, which together form user motivation. Each of these three are hypothesized to mediate the relationship between system characteristics (external variables Xs) and actual usage. A plethora of research has used the technology acceptance model and validated the psychological predictors of adoption. In later versions, attitude was swapped out for behavioral intentions because it had more explanatory power [50].

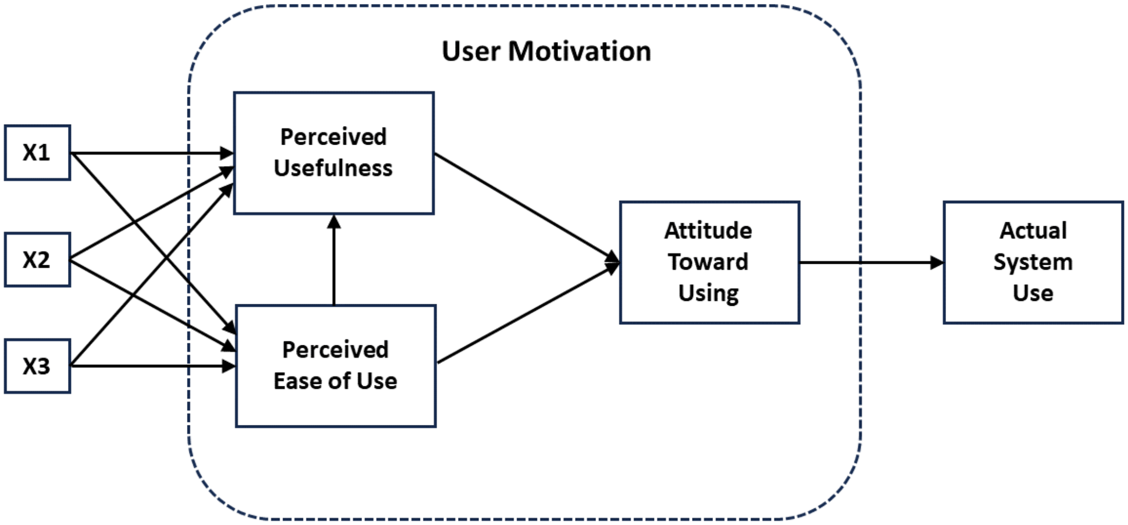


Figure 11. Technology acceptance model (image reproduced from [50]).

Other psychological variables such as user personality have been shown to increase likelihood of adoption [51]. According to [9] users fall into one of five personality-based groups that predicts how people or an industry will accept technological innovations. These are:

1. Innovators, who are willing to take risks and are the first ones to adopt an innovation
2. Early adopters, who adopt an innovation slower than innovators but quicker than other groups
3. Early majority, who adopt an innovation significantly after the innovators and early adopters but are still at or above average overall
4. Late majority, who adopt an innovation after the average time
5. Laggards, who are the last to adopt an innovation.

In the study, they asked 12 industry personnel in which phase they believed the nuclear power industry to currently be, with respect to adoption of AI technologies (in 2021). A majority indicated that they believed the nuclear power industry was early adopters of AI technology. This matches the commonly heard saying within the industry, in that there is a “race to be second” with new technologies. Typically, demonstrable and proven benefits must first be observed to move forward with the business case. The second most endorsed response was that the nuclear industry were innovators. However, some respondents indicated that there was variability between utilities, with some having used AI/ML for a decade while others are only getting started. Interestingly, predictive maintenance was named as a focus area. Taken together, individual and organization differences (personalities) should be considered with the adoption and implementation of AI/ML technologies in understanding which phase of adoption they may currently be in.

2.6 Summary

In this chapter we reviewed barriers to AI/ML adoption in nuclear power and demonstrate their interconnectedness. Nuclear power exists within a rich and storied history, with shifting public perceptions and political influences shaping investment in advanced nuclear technologies over the decades. AI-guided systems can support a better understanding of current plant status and upcoming maintenance requirements. Accurate, real-time predictions can optimize operator decision-making. AI is ubiquitous in other industries, including other energy sectors, and a key industry message today is “modernize, or get left behind” [52]. AI/ML is key to industry survival. The development, implementation, and sustainment of advanced AI-guided technologies will require many different types of expertise to maintain and regulate properly. These include technical, business case, regulatory and legal, policy, and HFE expertise, and must not be considered in isolation. AI demands bringing together a truly multidisciplinary team of experts and, to the best of our ability, a clear understanding of the broader societal implications.

3 DESCRIPTION OF THE SYSTEM AND FAULT MODES

This section focuses on describing the data and systems used for our research-developed solutions for some of the aforementioned barriers. The system chosen was a non-safety-related system at the Salem NPP, namely the CWS. It serves as the heat sink for the main steam turbine and associated auxiliaries. The CWS is designed to maximize steam power cycle efficiency while minimizing any adverse impact on the Delaware River [53]. The CWS consists of the following major equipment:

- Six vertical, motor-driven circulating pumps (or “circulators”), each with an associated trash rack and traveling screen at the pump intake to remove debris and marine life
- Main condenser (tube side only)
- Condenser waterbox air removal system
- Circulating water sampling system
- Screen wash system
- Necessary piping, valves, and instrumentation/controls to support system operation.

WBF 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 as the circulating water pump (CWP) intake comes directly from the river, resulting in a significant quantity of grass and debris. Primary symptoms of WBF 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 12 shows a schematic of the CWP and motor, including measurement locations.

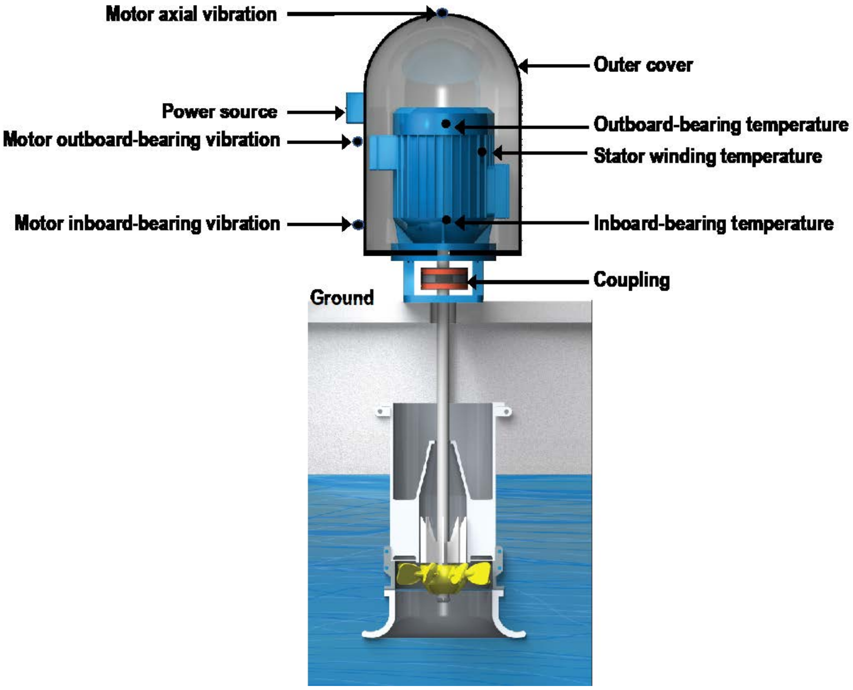
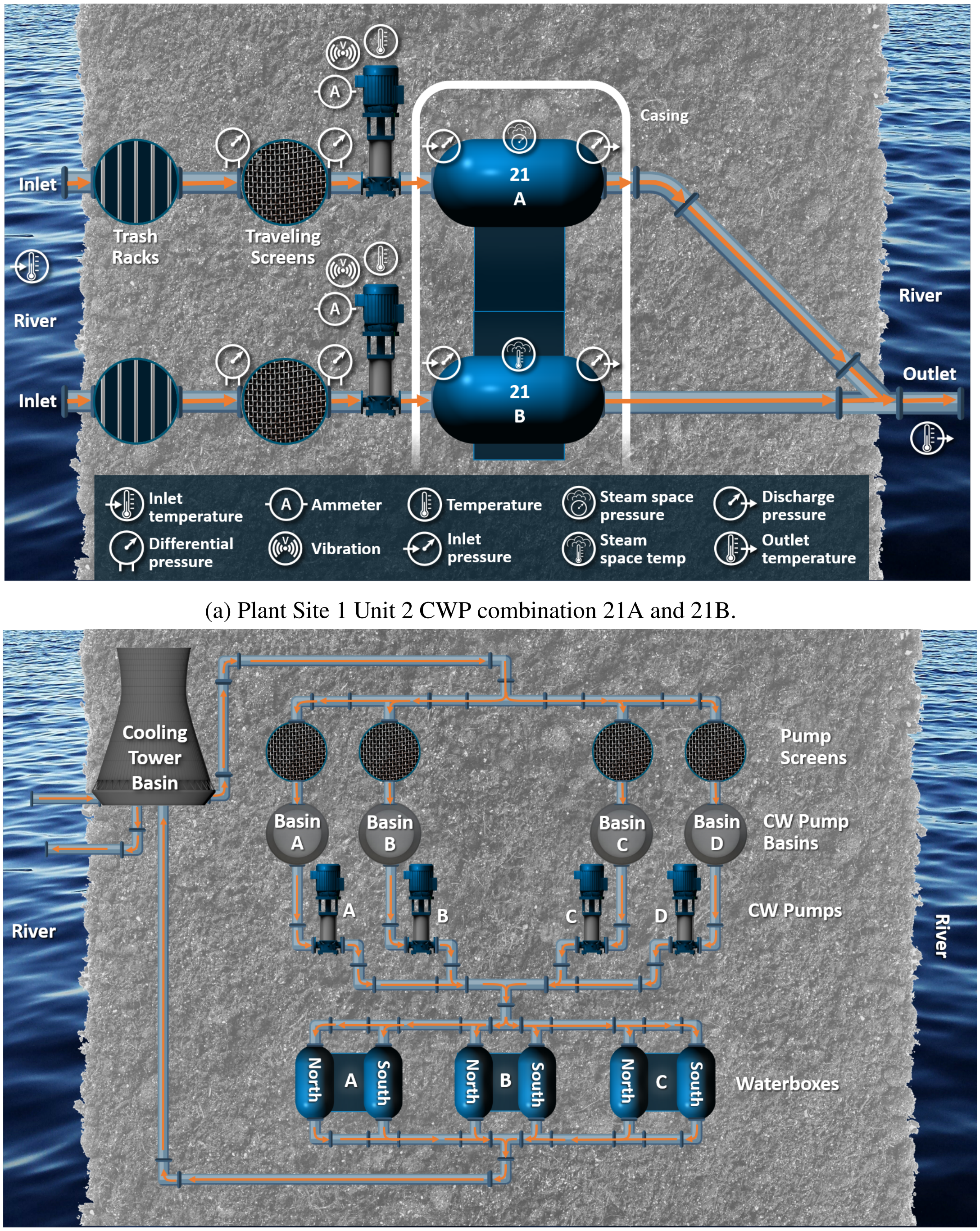

Motor axial vibration
Outer cover
Power source
Motor outboard-bearing vibration
Outboard-bearing temperature
Stator winding temperature
Motor inboard-bearing vibration
Inboard-bearing temperature
Coupling
Ground

Figure 12. Schematic representation of a CWP and motor, along with vibration and temperature measurement locations.

Plant Site 1 (a two-unit pressurized water reactor) features six circulators at each unit. Schematic representations of the main condensers for Plant Site 1, Unit 2 are shown in Figure 13(a). Each pair of waterboxes is named using the following convention: Unit #, Condenser #A, and Unit #, Condenser #B. Plant Site 2 (a single-unit boiling water reactor) has four circulators. A schematic representation of the Plant Site 2 CWS is shown in Figure 13(b). Several distinct differences can be seen when comparing the CWS schematics for Plant Sites 1 and 2, including that the water supply to the Plant Site 2 CWS comes from a cooling tower water basin, not directly from the river; the Plant Site 2 CWS does not have traveling screens, but each circulator has a single-pump screen to prevent debris transmission to the waterboxes; and the Plant Site 2 CWS has four circulators feeding six waterboxes via a common header, unlike the Plant Site 1 CWS, in which each waterbox had its own circulator.

A general functional description of the Plant Site 1 CWS, component integration, and design basis are found in [54]. This description is similar for the Plant Site 2 CWS, with minor differences in the integration as a result of previously highlighted changes in layout.



(a) Plant Site 1 Unit 2 CWP combination 21A and 21B.

(b) Plant Site 2.

Figure 13. Schematic representation of CWS.

The CWS has a major impact on the unit's gross load output (i.e., electricity production). There are also other systems in the plant that impact the unit's gross load output, but their impacts are minimal and hence they are not considered in this paper. Based on the number of CWPs operating at a given time, the unit's gross load can be labeled as full load, derate, trip, or outage (as shown in Figure 14). A derate is a percentage decrease in gross load due to the unavailability of a specified number of CWPs. A trip is a zero-power state when more than 50% of the CWPs are unavailable, leading to a forced shutdown. An outage is a planned

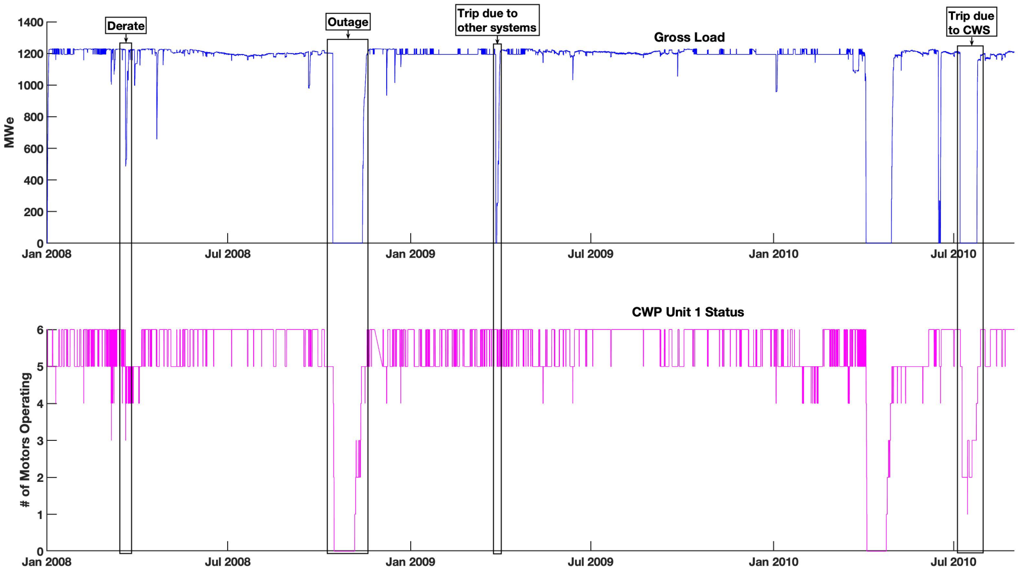


Figure 14. A sample of data showing the impact of availability of number of circulating water motors on the gross load.

shutdown when the reactor power is zero for an extended period of time (though usually less than a month) to complete scheduled fuel cycle maintenance.

The CWS process data for both units at Plant Site 1 are collected once per minute and stored in the plant's OSIsoft Process Information system. Due to file size restrictions, the process data for Plant Site 1 used in this work were downsampled to hourly data from 2009 to 2020. Similarly, the Plant Site 2 process data consist of hourly measurements spanning from January 1, 2010, to May 18, 2021. Table 1 details CWS-specific measurement types observed at both Plant Site 1 and 2.

Table 1. CWP-specific measurement types observed at Plant Site 1 and 2 NPPs.

Measurements	**Plant Site 1**	**Plant Site 2**
Timestamp	✓	✓
CWP status	✓	✓
Gross load	✓	✓
Differential temperature	✓	✓
Motor current	✓	×
Motor Stator temperature	✓	✓
Motor inboard bearing (MIB) temperature	✓	✓
Motor outboard bearing (MOB) temperature	✓	✓
Motor vibration (axial, outboard, inboard)	✓	×

The CWSs at both the plant sites experienced several types of faults in the time span for which the data were analyzed. These faults, as listed in Table 2, are infrequent, but the failure to diagnose them in a timely manner can result in unexpected downtime, derates, or trips, causing a drop in gross load that, in turn, leads to foregone revenue (i.e., lost opportunities to generate electricity and revenue) and additional maintenance costs. Based on the time period encompassing the data for analysis, some fault types resulted in multiple plant derates and trips (thus impacting plant generation), while others impacted plant generation only once or not at all.

For these diagnosed faults, relevant CWS process data, vibration sensor data, and work order data associated with the CWS were used to develop a condition-based monitoring solution. The CWS work order data [2] were used to create an approximate timeline of when faults occurred and were corrected, in addition to a timeline of PM activities. The fault timeline is particularly important for identifying possible fault features relevant to the fault modes listed in Table 2. ML models can be used to make such diagnoses based on fault signatures.

Table 2. CWS faults observed at Plant Site 1 and 2 NPPs.

Faults	**Plant Site 1**	**Plant Site 2**
WBF	✓	✓
CWP diffuser	✓	×
CWP bellmouth	✓	×
CWP shaft misalignment	✓	×
Clogging in air intake screens of the CW motors	✓	×
Moisture and salt contamination of CW motor windings	✓	×
CW motor oil (bearing oil) level (low)	✓	✓
Heating, ventilation, and air conditioning	×	✓
Inlet and outlet issue	×	✓
CWP screen clogging	×	✓
CW motor bearing	×	✓

This chapter focused on describing the data and systems used throughout the remainder of the report. Although the primary focus lies with WBF within the CWS, the knowledge gained from applying and explaining ML methods can be generalized to other systems.

4 EXPLAINABLE ARTIFICIAL INTELLIGENCE APPROACHES TOWARD TRUSTWORTHINESS

In this chapter, solutions for making ML more explainable and trustworthy are discussed, specifically as they relate to the CWS system and fault modes outlined in the previous chapter. This includes the ML's performance versus explainability trade-off, qualitative metrics to measure this, the value of new information when making a prediction, and novelty detection for knowing when an ML model may be extrapolating.

4.1 Performance versus Explainability Trade-off

As suggested in literature [55], there is an inherent tension between ML performance (e.g., predictive accuracy) and explainability, where the highest-performing methods (like deep learning) are the least explainable, and the most explainable (like decision trees) are the least accurate. It is further suggested in literature [56] that such a statement is only true when (1) the function to be approximated entails certain complexity; (2) the data available for study is greatly widespread among the world of suitable values for each variables; and (3) there is enough data to harness a complex model. Figure 15 shows the synthetic learning performance versus explainability trade-off for several categories of learning techniques.

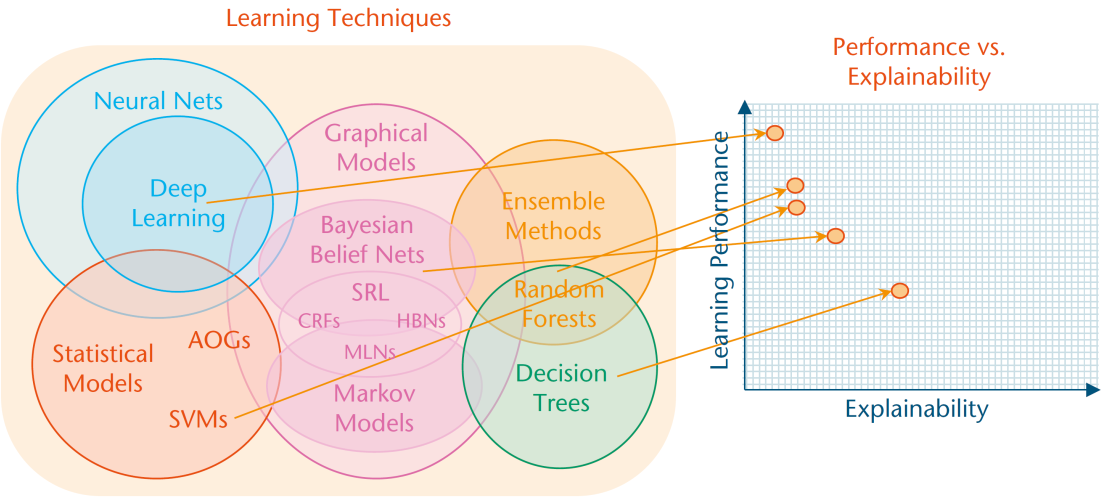


Figure 15. Synthetic learning performance versus explainability trade-off for several categories of learning techniques from [55].

Meanwhile, the literature [57] suggests that for cases with well-structured data and features, there is often no significant difference in performance between more complex classifier (deep learning, boosted decision trees) and much simpler classifier (logistic regression, decision lists) after preprocessing. The paper [57] also argues that the comparisons among different ML methods should not be performed on a static data set or training process since any formal training process that extracts knowledge from data requires an iterative process. More iterations lead to a model with better training and more interpretable results. In cases with more iterations, the accuracy/explainability trade-off is reversed: instead of having less explainability for a more accurate model, more explainability leads to better overall accuracy.

To test such a hypothesis on performance-explainability trade-off, this section selects six ML methods, including a Feedforward Neural Network (FNN) model with complex architectures, Extreme Gradient Boosting (XGBoost) [58] based on gradient boosting algorithms, k-nearest neighbors (KNN), decision tree, Bayesian classification, and a logistic linear regression model.

Among the faults of interest (Section 3) examined for the CWS, only four fault types had multiple instances that resulted in CWP shutdowns. The multiple occurrences of these faults allow for development and testing of diagnostic models, strengthening the fault signature. WBF caused numerous instances of CWP shutdowns. Even though this fault is not a pump or motor fault, it is a system fault that may show symptoms affecting pump performance. The WBF fault also occurred numerous times, allowing for development and testing of diagnostic models. Fault types with only a single instance of causing a CWP shutdown provided limited information for developing a fault signature and training ML. Table 3 lists the number of data instances for six faults and normal conditions.

As a result, this study focuses on a diagnosis model for WBF. A total number of 220,702 data instances from plant site 1 are used, including 46,289 instances labeled as WBF and 174,413 instances labeled healthy conditions. Table 4 lists the number of data instances for all six groups of CWP with healthy and WBF labels.

Table 3. Number of data instances for healthy conditions and each fault modes.

Modes	**Number of data instances**
Healthy	174,413
Air intake	855
Bellmouth fail	417
CWP diffuser	502
Misalignment	5,237
WBF	46,286

Table 4. Number of data instances for healthy conditions and each fault mode.

Group number	**Number of healthy instances**	**Number of WBF instances**
1 (CWP11A & CWP11B)	13,455	2,691
2 (CWP12A & CWP12B)	19,120	3,824
3 (CWP13A & CWP13B)	19,120	3,824
4 (CWP21A & CWP21B)	5,825	1,165
5 (CWP22A & CWP22B)	7,250	1,450
6 (CWP23A & CWP23B)	5,795	1,159

The healthy data are about four times as much as WBF data. For each instance, all features listed in Table 1 (except for motor vibration) are included in the training and testing process. The motor vibration feature is excluded because vibration sensors were recently deployed and cover a very short time period, which significantly reduces the number of data instances for training and testing of diagnosis models. In addition, Week of the Year (WoY), and motor and pump age are added to account for seasonal and long-term effects respectively.

To evaluate their performance, different combinations of training and testing data are selected, including stratified random sampling and stratified group-based sampling. The goal is to avoid biases in selecting either a non-representative or lack of diverse data for performance evaluations [59]. For stratified random sampling, a fixed percentage of data instances is selected from data instances labeled as WBF and healthy, respectively. However, simple random sampling strategies are known to have issues in representing minority subgroups with populations not fully represented in samples [60]. Moreover, the generalization capabilities of diagnosis models can hardly be evaluated with random sampling strategies. The generalization capabilities refer to model's ability to adapt properly to new and previously unseen data [61]. Therefore, the group-based sampling strategy is utilized to evaluate the model's performance in predicting unseen data instances from different CWP groups. For stratified group-based sampling, data instances from different combinations of groups are used as the training and testing data, respectively. As shown in Figure 13a, one CWP combination/group contains two pumps, and there are six groups in total at the Salem plant site.

Moreover, as shown in Table 4, there are more data instances with healthy status than instances with WBF, and such a data imbalance is known to affect the performance of ML methods [62], which thus will affect the performance explainability trade-off. As a result, this study also investigates data balance issues by truncating healthy data instances into the same amount, twice, three times, and five times as much as the WBF instances. Overall, based on different sampling strategies and levels of data unbalance, the performance is evaluated by diagnosing testing data labels. The performance of each ML method is compared to determine if the performance-explainability trade-off is a valid hypothesis. Table 5 summarizes three case studies for testing the hypothesis.

Table 5. Case studies with different sampling strategies for the testing performance-explainability trade-off hypothesis.

Sampling Strategies	Description
Stratified random sampling	For data with healthy and waterbox-fouling labels, 80% of data instances (20,000 data points) are randomly selected as training data and the rest, 20%, (5,000 data points) is selected as testing data, respectively. For both training and testing data, there are the same amount of data instances (12,000 data points) with healthy and WBF labels. Each data instance contains seven features.
Stratified group-based sampling	For each group, the number of healthy data is truncated to the same amount of WBF (as in Table 4). For a designated number of groups, ranging from 1–5, all possible of combinations of CWP groups, based on Table 4, are selected as the training data, while the rest of the groups are used as the testing data in a cross-validation manner. Each data instance contains seven features.
Unbalanced data	For each group, the number of healthy data is truncated to one times, two times, three times, and five times to the WBF. Based on Table 4, all combinations of 5 groups of CWP are selected as the training data, while the remaining 1 group is used as the testing data in a cross-validation manner. Each data instance contains 7 features.

To account for the iterative process in model training, the hyperparameters are sampled for each ML method. Hyperparameters refer to parameters whose values control the model training/learning process and determine the values of model parameters that a learning algorithm ends. For each ML method, important hyperparameters are optimized for minimizing/maximizing a cost function based on the model performance. The objective is to (1) avoid model errors and biases in setting up the training problems and (2) identify the uncertainty of ML models in extracting knowledge from the data through an iterative process. Table 6 lists high-impact hyperparameters of all six models. In this work, 1000 hyperparameters are drawn from the candidate values using the Tree of Parzen Estimators (TPE). The TPE algorithm uses a truncated Gaussian mixture model (GMM) [63] to fit the high-dimension correlations between the hyperparameters and the model performance. A GMM is a weighted sum of non-parametric component Gaussian densities based on continuous-valued data vector (i.e., the output of cost function). The sum of weights are required to be 1, and the non-parametric densities will be updated by observations, representing a learning algorithm that produces a variety of densities over the configuration space [64].

This work uses recall rates for predicting WBF as the performance measure metrics, and the TPE algorithm will maximize the expected improvement based on the performance results from sampling histories for optimizing the hyperparameters drawing from new iterations. Recall rates refer to the fraction of true WBF instances over all WBF labels that are predicted by ML methods. The goal is to reduce false alarm

Table 6. List of ML methods and hyperparameters.

ML Methods and Hyperparameters	Class for Drawing Candidate Values
Linear Model	
Precision of the regression solution	uniform distribution: [1e-6, 1e-2]
Regularization strength	uniform distribution: [0, 1]
Bayesian Network	
Portion of features for variance calculations	uniform distribution: [1e-9, 1e-4]
K–Nearest Neighbor	
Number of neighbors for majority vote	uniform distribution: [1, 10]
Algorithms for nearest neighbor computations	brutal force, ball tree, k-dimension tree
Decision Tree	
Functions for measuring the quality of split	gini, entropy, log loss
Maximum depth of the tree	uniform distribution: [1, 20]
Minimum number of samples required to split an internal node	uniform distribution: [0.1, 0.5]
Number of features to consider when looking for the best split	uniform distribution: [1, 7]
XGBoost	
Number of gradient boosted trees	uniform distribution: [10, 250]
Maximum tree depth for base learners	uniform distribution: [10, 45]
L1 and L2 regularization terms on weights	uniform distribution: [0, 1]
FNN	
Number of layers	uniform distribution: [1, 5]
Number of neurons per layers	uniform distribution: [10, 50]
Dropout ratio	uniform distribution: [0.01, 0.5]
Validation patience	uniform distribution: [5, 20]
Batch size	uniform distribution: [100, 1000]

rates and improve economic viability of existing NPP by accurately predicting all WBF instances. For ML methods that require early stopping, including XGBoost and FNN, 20% data are randomly drawn from the training data to serve as the validation dataset.

Figure 16 compares the feature distributions from training and testing data based on random sampling, and very few differences can be visualized. However, as shown in Figure 17, the feature distributions of testing data are very different from the distributions of training data. As a result, the performance of the ML-based diagnosis model based on six different ML methods, including FNN, XGBoost, KNN, decision tree, Bayesian network, and linear model is affected as shown by the recall rates in predicting the WBF. The predictions are made on testing data that are selected randomly and based on the groups, respectively.

Figure 18 plots the distribution of model performance in predicting the testing data with random sampling. Median recall rates are also shown to avoid skewness. Gaussian Bayes and linear ridge classifier have slightly lower median recall rates than neural network (NN), XGBoost, KNN, or decision tree. Despite better performance, recall rates of NN and decision tree have larger variance. It suggests that the prediction results from NN and decision tree are more likely to be affected by hyperparameter selection. Similar trends in recall rates are observed for all ML methods in predicting the training data.

Figure 19 compares the distribution of model performance with one group selected as the testing data during the training and testing stages. The performance of KNN in predicting the testing data is lower than predicting the training data since KNN only stores training data instead of learning and extrapolating from the data. The variances of all models increase (i.e., the performance is reduced) when they are used to predict

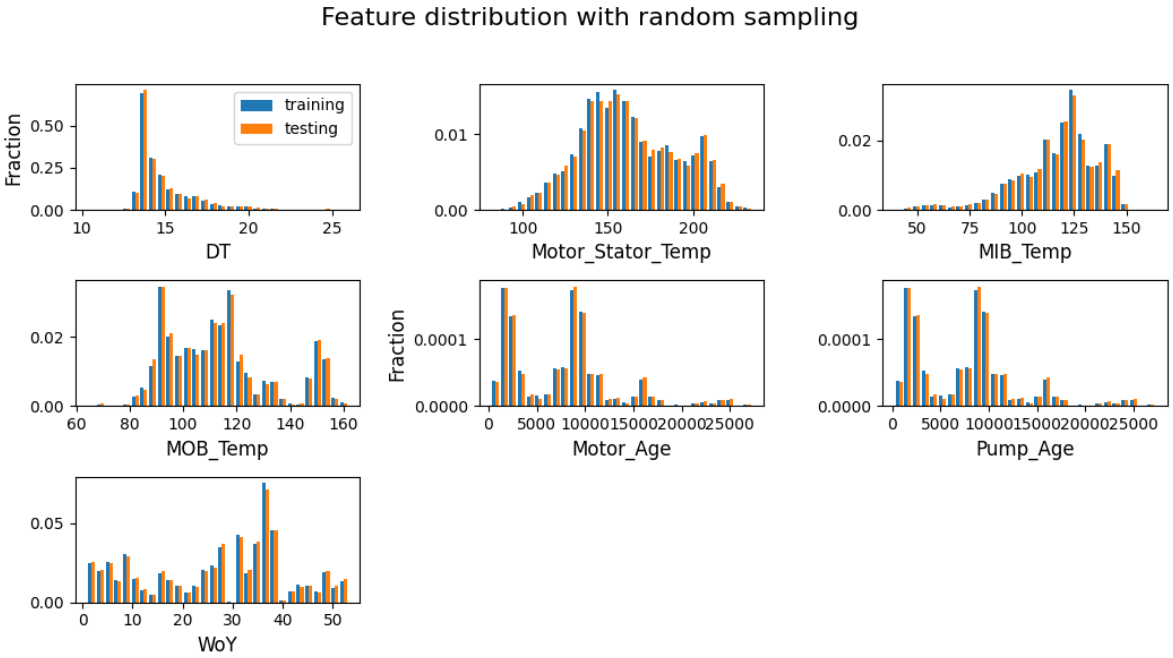


Figure 16. Comparison of training and testing data distributions with random sampling.

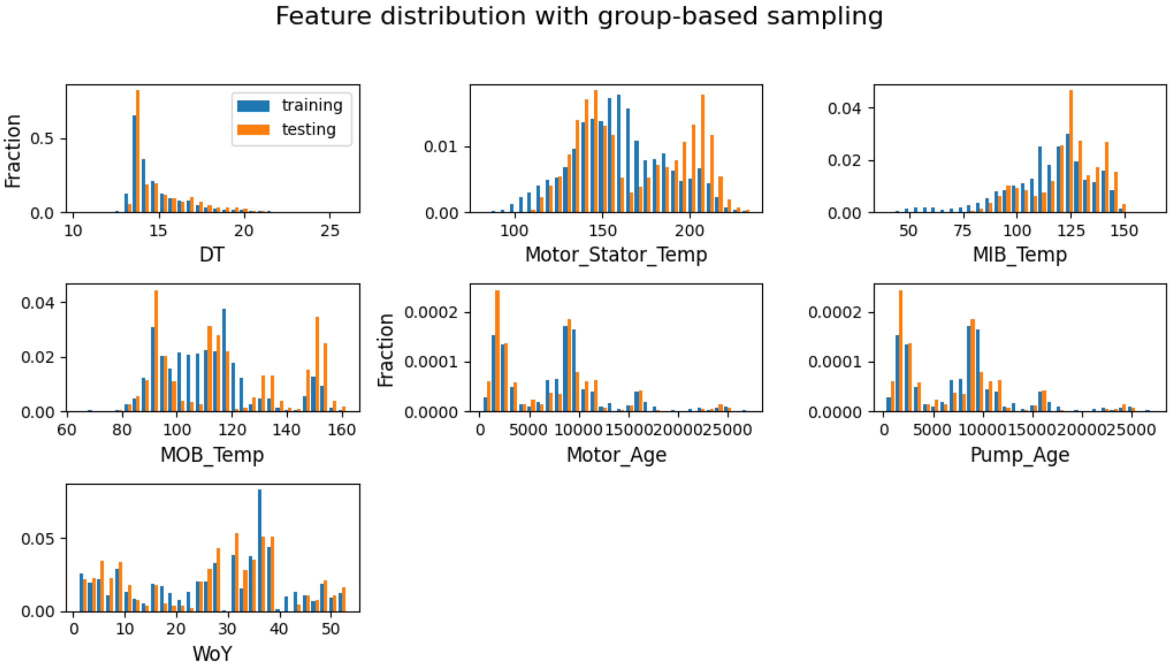


Figure 17. Comparison of training and testing data distributions with random sampling.

unseen data, which are outside the distribution of training data. NN and decision tree tend to have larger variances than other models.

Table 7 lists the ML model performance measured by recall rates with a different number of groups compared to the testing data. For each number of groups, all combinations of groups are used as the training data, while the rest of the groups are used as the testing data. The medians of recall rates are determined

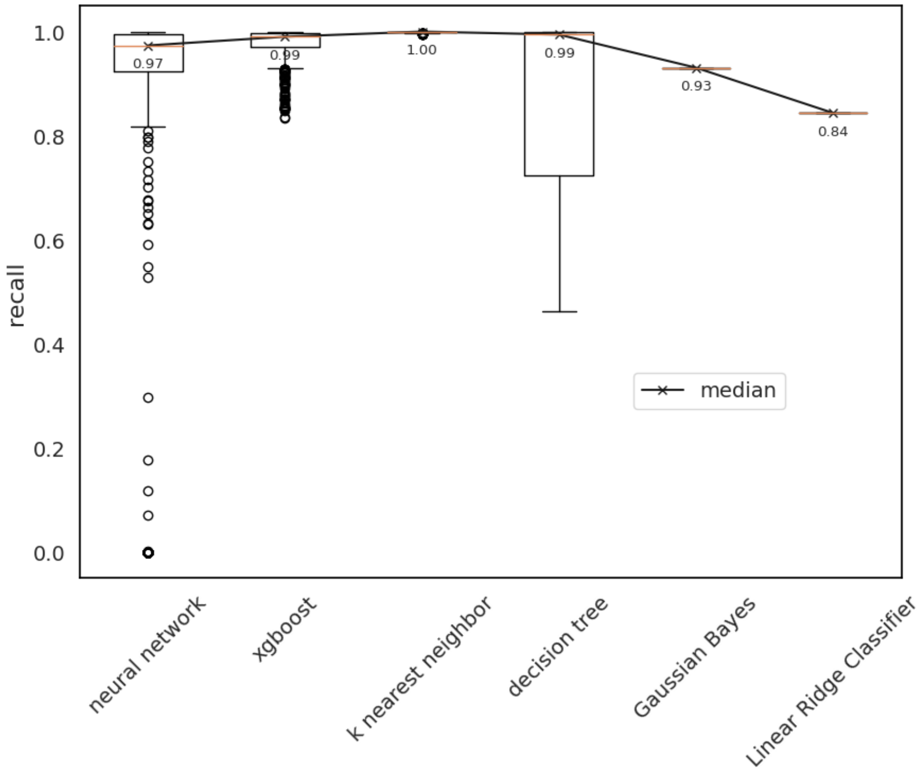
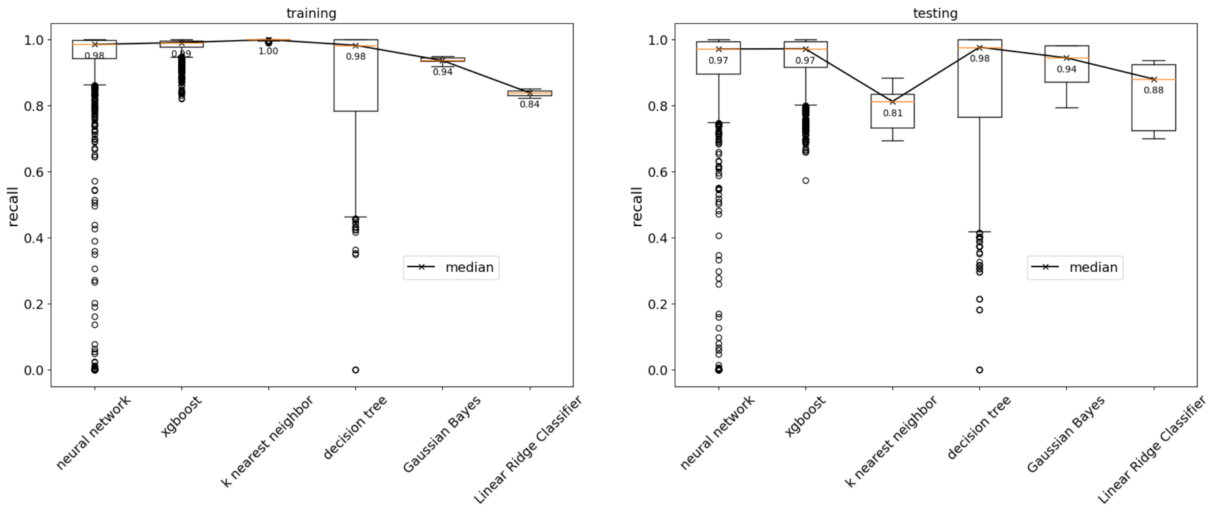


Figure 18. Comparison of training and testing data distributions with random sampling.



(a) (b)

Figure 19. ML model performance for group-based sampling measured by recall rates in predicting (a) training data and (b) testing data.

by aggregating testing results from all groups. Again, the goal is to reduce false alarm rates and improve economic viability of existing NPP by accurately predicting all waterbox-fouling instances, so recall was the performance metric of choice. No significant differences are observed among testing results with a different number of groups.

Figure 20 further shows the relationship between number of tunable parameters in a FNN and the median

Table 7. Median of recall rates with different number of groups as testing.

ML Methods	**Number of Groups Used as Testing Data**				
	1	2	3	4	5
NN	0.97	0.97	0.96	0.96	0.94
XGBoost	0.97	0.97	0.97	0.97	0.94
KNN	0.81	0.88	0.87	0.86	0.84
Decision Tree	0.98	0.99	0.98	0.98	0.95
Gaussian Bayes	0.94	0.93	0.94	0.95	0.97
Linear Ridge Classifier	0.88	0.81	0.82	0.83	0.81

of recall rates in predicting WBF in the testing data. As an indicator of model complexity, more tunable parameters in NN suggest more complex and less explainable models. However, no significant improvements are observed when more complex NN are trained and used for predicting unseen data points outside the training distributions.

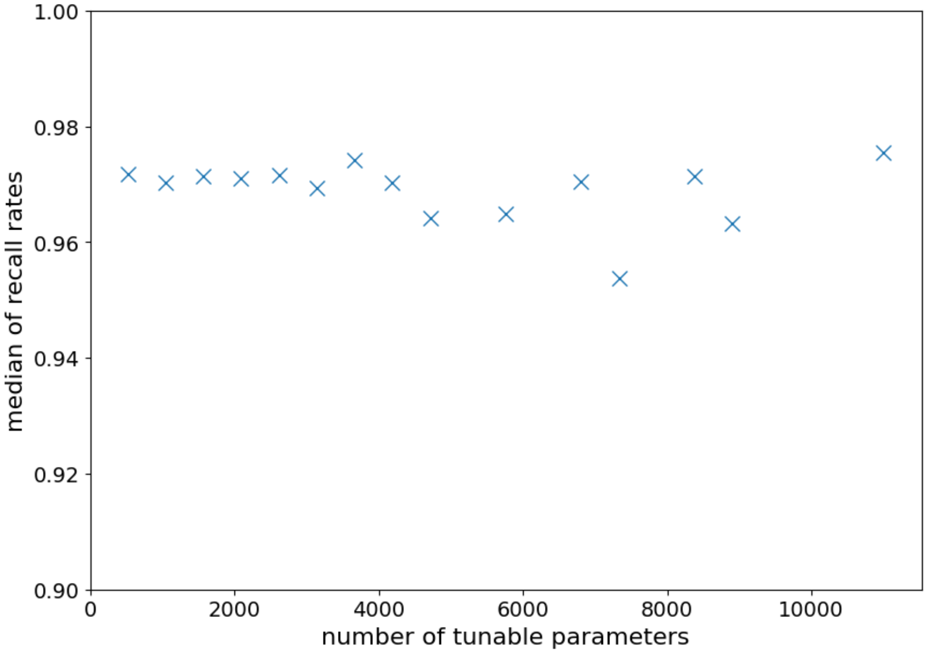


Figure 20. Relationship between number of tunable parameters in NN and the median of recall rates.

Since the performance-explainability trade-off based on the literature is only true when the data are not well processed and well structured, this study further compares performance of different ML methods with unbalanced training data. Specifically, the training data with label of 'healthy' appear twice as often as points with 'WBF' label. Operational data from one of six groups are used as testing data. Figure 21 shows the distribution of model performance with healthy data twice as much as WBF data.

Figure 22 further shows the distribution of model performance with healthy data is five times more than WBF data. Compared to Figure 21, performance of all models had lower performance levels as the training data become more unbalanced. Specifically, the median recall rates of the decision tree model drop to 0, meaning decision trees tend to label all data as healthy. The median recall rates of Gaussian and linear classifier drop more significantly than neural networks and XGBoost models. Moreover, when data are very unbalanced, complex models like FNN and XGBoost perform better than Gaussian Bayes and linear classifier with simple model forms. Table 8 summarizes the median recall rates of all model approaches with different degrees of unbalanced training data, which are measured by the relative ratios between the number of healthy data and WBF data.

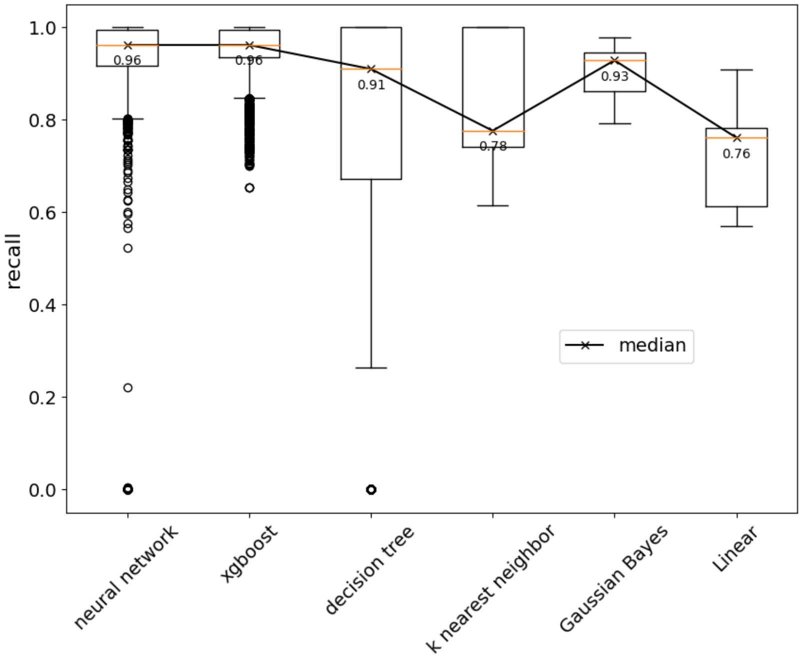


Figure 21. Distribution of model performance trained by healthy data twice as WBF data.

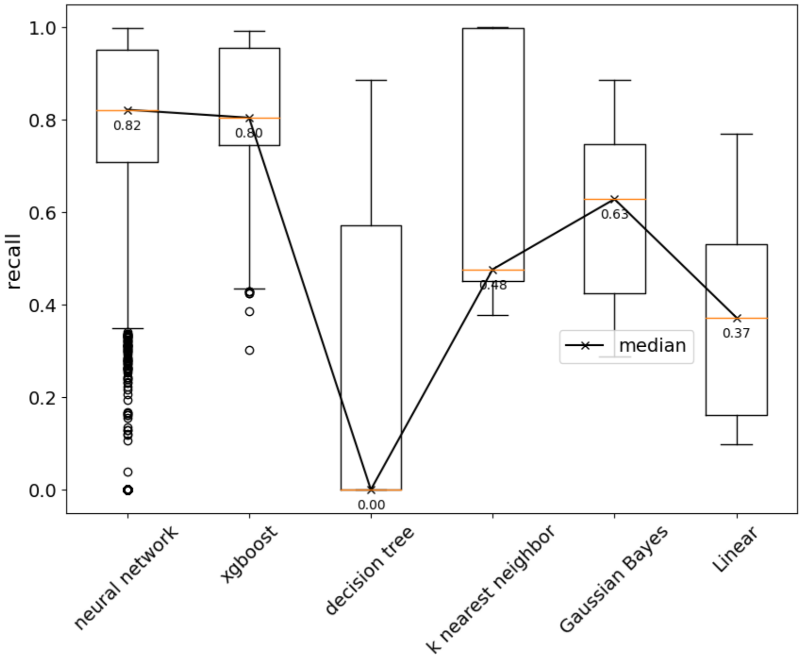


Figure 22. Distribution of model performance trained by healthy data five times as WBF data.

In conclusion, this work selects six ML methods, including complex models like FNN and XGBoost and simple models like Gaussian Bayes and linear classifier. After training each model with operation data from six groups of PSEG-owned NPP CWPs, the distribution and median of model performance for all ML methods are similar when the training data are processed and balanced. Such similarities are observed when

Table 8. Median recall rates of all ML methods with different degrees of data unbalance.

MLmethods	Number of healthy data with respect to WBF data			
	1	2	3	5
NN	0.97	0.96	0.97	0.82
XGBoost	0.97	0.96	0.92	0.80
KNN	0.81	0.91	0.71	0.48
Decision Tree	0.98	0.78	0.66	0.00
Gaussian Bayes	0.94	0.93	0.84	0.63
Linear Ridge Classifier	0.88	0.71	0.80	0.37

testing data are selected randomly or when different numbers of groups are used as testing data. Meanwhile, to test the impacts of balanced and unbalanced data, this work selected training data with different ratios in labels. Compared to simple models like Gaussian Bayes and linear classifiers, FNN and XGBoost have better performance when data with healthy labels that contain the same, twice, three times, and five times as much as data compared with WBF. As a result, complex models could have better performance than simple models when data are not well structured or well processed. For well-processed data with balanced data labels, the performance is comparable with simple or complex models.

4.2 Qualitative Attributes

In addition to a working definition of ML explainability, the essential features of the explanatory demands of explainable AI need to be investigated to better understand technical and delivery dimensions of AI explainability. Such an investigation involves making explicit how a particular set of attributes can play the role of evidence in supporting the conclusion reached by AI/ML models. The process of determining and assembling attributes should give decision-makers the rationale behind that AI/ML result as if it had been produced by a reasoning, evidence-based, and inference-making person. This section reviews qualitative attributes that can be used to support the explainability and thus the confident use of AI/ML tools.

It has been suggested in [65] that explaining an algorithmic model's decision should make explicit how the particular set of factors, which affect the outcome of ML models, can play the role of evidence in supporting the conclusion reached. To build a reasoning process toward this explanation-giving task, human-scale reasoning and interpreting includes four aspects:

- Logic and techniques in applying the basic principles of validity that lie behind and give form to sound thinking. The goal is to provide a step-wise application of procedures and rules that comprise the formal framework of the algorithmic system. Such a goal is supported by evidence from modeling, approximating, and simplifying the most complex and high dimensional computations.
- Semantics in gaining an understanding of how and why things work the way they do and what they mean. The goal is to understand the meaning of procedures and rules in terms of their purpose in the input-output mapping operation of the system. Such a goal is supported by reasoning the relation of the predictor and response variables.
- Social understanding of practices, beliefs, and intentions in classifying the content of interpersonal relations, societal norms, and individual objectives.
- Moral justification in making sense of what should be considered right and wrong in everyday activities and choices.

In addition to the high-level attributes, it is suggested in [56] that XAI techniques can be classified into transparent models and post-hoc explainability. Transparent models convey degrees of explainability by themselves. Models belonging to this category can be explained in terms of simulatability, decomposability, and algorithmic transparency.

- Simulatability refers to the ability of a model being simulated or thought about strictly by a human, hence complexity takes a dominant place in this class.
- Decomposability refers to the ability to explain each of the parts of a model (input, parameter, and calculation). It can also be considered as intelligibility [66].
- Algorithmic transparency deals with the ability of the user to understand the process followed by the model to produce any given output from its input data.

Post-hoc explainability aims to enhance the interpretability of black-box models that are not readily interpretable by design. Specific techniques include text explanations, visual explanation, local explanations, explanations by example, explanations by simplification, and feature relevance explanation methods. Based on this classification, evidence on model explainability for different ML approaches is shown in Table 9:

Table 9. Evidence about model explainability based on four attributes. Adopted from reference [56].

Model	Model Transparency			Post-hoc Analysis
	Simulatability	**Decomposability**	**Algorithmic Transparency**	
Linear/Logistic regression	Predictors and interactions among them are human readable.	Variables are readable, but the number of interactions and predictors can be decomposed for better explanations.	Variables and interactions need to be analyzed with mathematical tools and metrics.	Not needed
Bayesian Model	Statistical relationships are modeled among variables and the variables themselves should be directly understandable by the target audience.	Statistical relationships involve many variables, which must be decomposed in marginals for analysis.	Statistical relationships cannot be interpreted even if already decomposed, and predictors need to be analyzed with mathematical tools and metrics.	Not needed
Decision Tree	Predictors and interactions among them are human readable.	Rules are readable, and the number of interactions among variables do not alter data.	Human readable rules and a direct understanding of the prediction process.	Not needed
KNN	Predictors and interactions among them are human readable.	The number of variables is too large and needs decomposition for analysis.	Mathematical tools and metrics are needed to analyze the model.	Not needed
XGBoost	Predictors and interactions are not readable.	Universal decomposition techniques are not available.	Universal tools and metrics are not available.	Needed: usually model simplification, local explanations, or feature relevance techniques
Multi-layer neural network	Predictors and interactions are not readable.	Universal decomposition techniques are not available.	Universal tools and metrics are not available.	Needed: usually model simplification, local explanations, or feature relevance techniques

As a result, the explainability of ML methods can be expressed based on qualitative attributes and supporting evidence. Figure 23 shows explainability measures supported by qualitative attributes and corresponding evidence regarding simulatability, decomposability, algorithm transparency, feature contributions/relevance, and simple model approximations. Moreover, explainability of simple model approximation can be measured based on the same structures and corresponding evidence.

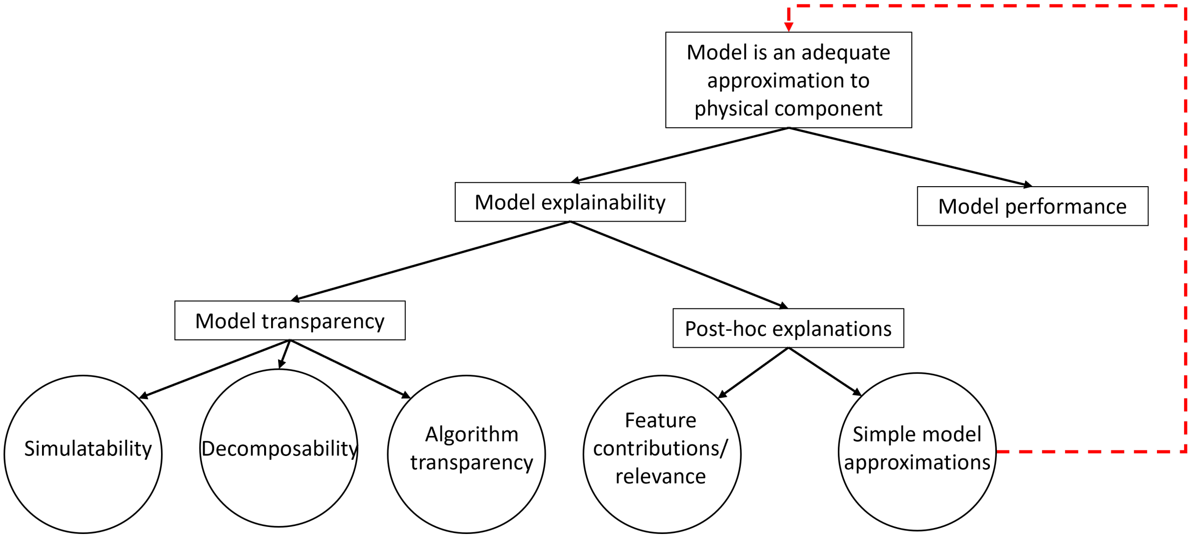


Figure 23. Explainability measures supported by qualitative attributes and corresponding evidence.

A user interface can also be built based on the explainability measures, which include evidence of each attribute mentioned above (as shown in Figure 24).

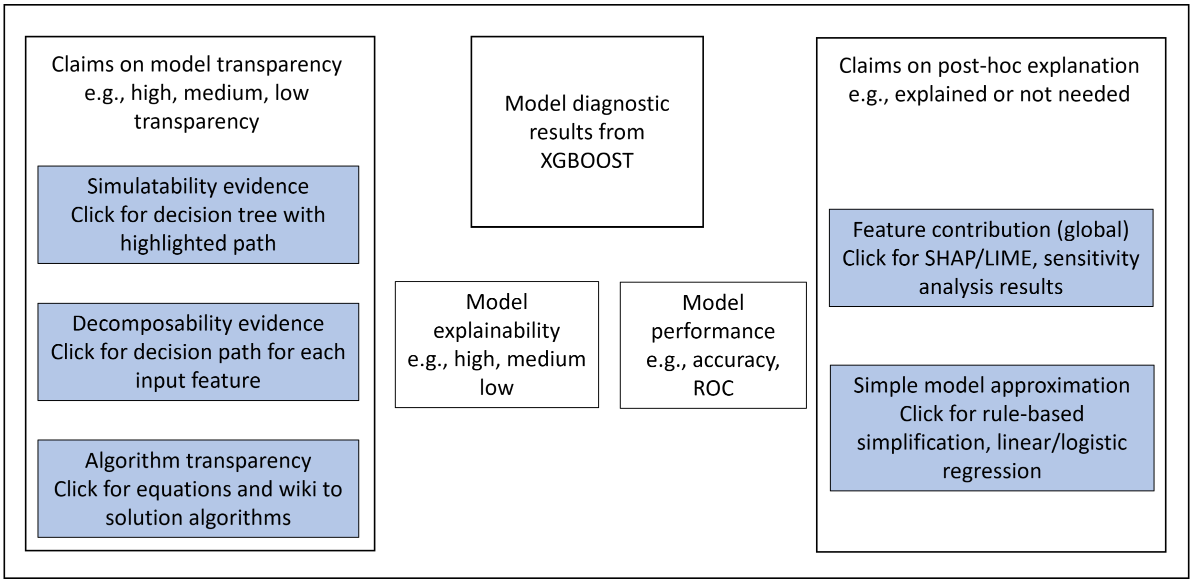

Claims on model transparency e.g., high, medium, low transparency
Simulatability evidence Click for decision tree with highlighted path
Decomposability evidence Click for decision path for each input feature
Algorithm transparency Click for equations and wiki to solution algorithms
Model diagnostic results from XGBOOST
Model explainability e.g., high, medium low
Model performance e.g., accuracy, ROC
Claims on post-hoc explanation e.g., explained or not needed
Feature contribution (global) Click for SHAP/LIME, sensitivity analysis results
Simple model approximation Click for rule-based simplification, linear/logistic regression

Figure 24. Explainability measures supported by qualitative attributes and corresponding evidence.

4.3 Value of New Information

The question of how much information is needed to inform a decision is central to applied ecology. Information from system or component monitoring is often vital to effective decision making. However, many decisions are made based on inadequate information [67], and monitoring can be costly and time-consuming. Therefore, it is important to choose monitoring strategies to gather only information that is necessary to make an effective decision [68]. The theory of the Value of Information (VoI) is a decision analytic method for quantifying the benefit of acquiring additional information to support such analyses [69]. The theory arises from considering jointly the probabilistic and economic factors that affect decisions. According to the VoI theory, the expected value $\mathbb{E}$ of a given action a_i under uncertainty can be calculated as:

$$\mathbb{E}[V(a_i, s)] = \sum_{s \in S} V(a_i, s) \cdot P_s \,. \tag{1}$$

where $V(a_i, s)$ is the value of an action for a state s; P_s is the prior probability of a state s; A represents a set of all possible actions for a given state s; This is the sum of all possible values for action a_i for all states s of the target units, each of which is weighted by its respective probability of the state s being true. As a result, the best action is to maximize the expected values from Equation 1:

$$EV_{uncertainty} = \max_{a_i \in A} \mathbb{E}_s[V(a_i, s)] \,. \tag{2}$$

The same equation can also be formulated as a decision problem by introducing a binary decision variable x_i representing whether action a_i is implemented:

$$EV_{uncertainty} = \max_{x_i} \sum_{i=1}^{|A|} x_i \mathbb{E}_s[V(a_i, s)] \,. \tag{3}$$

Monitoring will typically improve upon current information and increase the expected value of the best management action. Specifically, monitoring will the change decision maker's belief about the probability of each state s being true. In VoI, probabilities for each state s being true can be estimated based on each possible monitoring result using Bayes Theorem [70]. As a result, the expected value of the best management action when information from monitoring y can be described as:

$$EV_{monitoring} = \mathbb{E}\max_{a_i \in A} \mathbb{E}_{|y}[V(a_i, s)] \,. \tag{4}$$

Accounting for the decision variables x_i, the expected value for the best action for each monitoring result y becomes:

$$EV_{monitoring} = \max_{x_i^y} \sum_{y \in Y} p(y) \sum_{i=1}^{|A|} x_i^y \sum_{s \in S} V(a_i, s) p(s|y) \,. \tag{5}$$

where x_i^y identifies action a_i to implement for each possible monitoring result y, weighted by probability of obtaining the monitoring results. The expected value of monitoring information, also known as the expected value of sampling information (EVSI) is the difference between the expected value of the best action after monitoring and the expected value of the best management action under uncertainty before monitoring:

$$EVSI = EV_{monitoring} - EV_{uncertainty} \,. \tag{6}$$

For the maintenance based on diagnosis results, when the diagnosis results are uncertainty, the possible states, prior probabilities, and expected values from Equation 1 and 2 can be described using a decision tree as shown in Figure 25

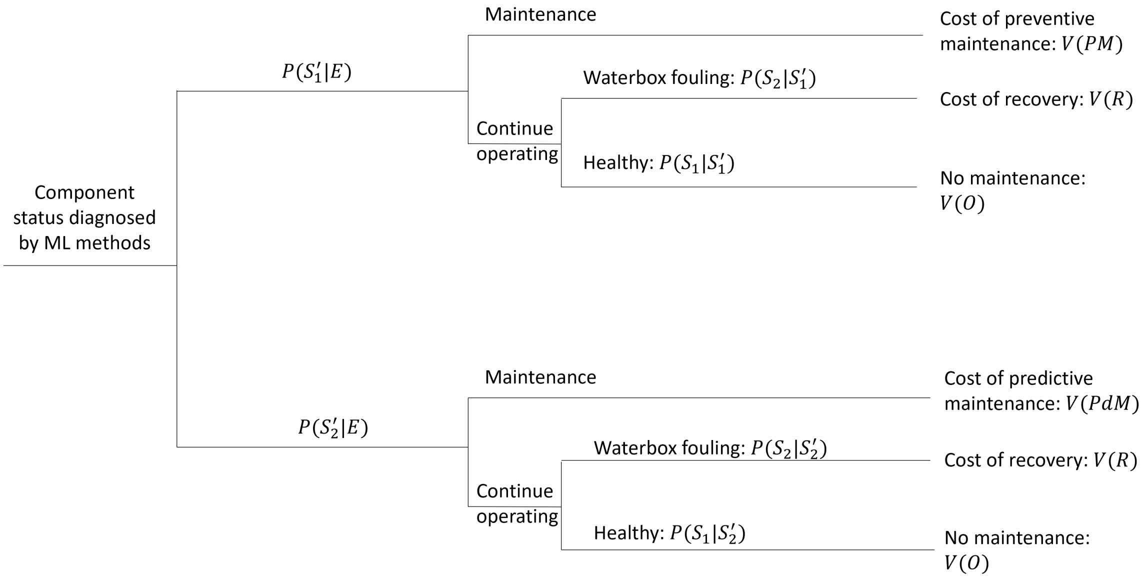


Figure 25. Decision tree model for maintenance based on diagnosis results. All possible states, prior probabilities, and expected values are included.

where E refers to evidence from an ML-based diagnosis model, operating histories, user interface, and additional information from monitoring. S_1' and S_2' are predicted states from the ML-based diagnosis. S_1 and S_2 are actual states of the component. $P(S_1'|E)$ and $P(S_2'|E)$ are the diagnosis model's confidence in equipment status given the evidence. In this example, S_1' and S_1 refer to a healthy CWP, while S_2' and S_2 refer to a CWP subject to WBF. As a result, the diagnosis model could suggest a healthy CWP based on evidence from sensor readings and monitoring activities, and the decision maker could choose a PM strategy and perform maintenance immediately. Otherwise, they could decide to continue operating the CWP, and the decision-maker could either do nothing or end up needing to recover from the WBF fault because the diagnosis model failed to detect fouling. Meanwhile, the diagnosis model could suggest a WBF will happen, and decision-maker can choose either to follow the PdM strategy or to neglect the warning and continue operation. If the decision-maker decides to continue operating, they would either have to recover from WBF or end up not needing any maintenance. For each decision path, costs are assigned for possible maintenance strategies, including the cost of PM $V(PM)$, cost of recovery $V(R)$, cost of PdM $V(PdM)$, and cost of no maintenance $V(O)$. This study assumes that (1) the cost of recovery from WBF is the largest and (2) that the cost of PM is larger than PdM. To demonstrate the VoI concept, this study assigns synthetic values for all conditional probabilities and cost values (Table 10). Compared to scenarios with no new monitoring information, the new monitoring information improves the confidence from the diagnosis model that the CWP is in a healthy condition. The differences between the expected value of costs yield 127.1 as the VoI. In this scenario, new monitoring reduces the expected costs by 15%.

4.4 Data Novelty

Another important aspect of model trustworthiness is whether the current data are consistent with or close to the training data. Caution must be taken when extrapolating with any model, ML-based or not.

Table 10. Synthetic values of conditional probability and values of maintenance costs.

	No monitoring information	New information from monitoring
$P(S_1\|E)$: Confidence in healthy diagnosis	40%	80%
$P(S_1\|E)$: Confidence in WBF diagnosis	60%	20%
$P(S_2\|S_1\prime)$: Confidence in false negative	20%	10%
$P(S_1\|S_1\prime)$: Confidence in true negative	80%	90%
$P(S_2\|S_2\prime)$: Confidence in true positive	80%	95%
$P(S_1\|S_2\prime)$: Confidence in false positive	20%	5%
$V(PM)$: Cost of PM	500	
$V(PdM)$: Cost of PdM	100	
$V(R)$: Cost of recovery	1000	
$V(O)$: Cost of no maintenance	10	
Expected value of costs	824.4	697.3

In general, extrapolation is more acceptable when there are physical reasons to believe the model form is correct and that it should apply in operating regimes not represented by the training data. Using extrapolation with purely data-driven models is, in general, a perilous course of action, particularly in situations where actions—such as issuing a work order—are dictated by the prediction outcomes of classification. There are multiple ways to identify situations in which the current data significantly deviate from the training data, including the use of novelty detection, distance metrics, and convex hull analysis. Presenting users with a visual indication that a current data point is consistent with the training data will boost their confidence in the model's output.

Novelty detection is commonly used in ML to determine whether a new data point is an outlier relative to the training data. Novelty detection is essentially a binary classifier that classifies a new data point as either consistent or inconsistent with the training data. There are many techniques for novelty detection, which can be classified into five broad categories: probabilistic, distance-based, reconstruction-based, domain-based, and information-theoretic [71]. Similarly, an analysis can be used to determine if the addition of the current data point to the training data set changes the convex hull for the data set (in two dimensions, the convex hull of the set is the smallest convex polygon that contains all the points) [72]. Although these methods provide an excellent way to identify data that are somehow different than the test data, they do not provide any type of intuitive explanation for the decision (e.g., Motor inboard bearing (MIB) temperature is too high).

A more intuitive approach leverages distance metrics to calculate how far the current data point lies from the training dataset. In two dimensions, this could be simplified by presenting a scatter plot showing where the data point lies relative to the training data. Because ML datasets tend to be multidimensional, distance metrics are useful for identifying situations where the data point under consideration is outside the training data range. For numeric data, the most common distance metrics are the Euclidean, Chebyshev, Manhattan, and Minkowski metrics [73].

4.4.1 Distance Metrics

The Euclidean distance is perhaps the most ubiquitous distance metric used. Given two data points in n-dimensional space, $\vec{x} = (x_1, x_2, ..., x_n)$ and $\vec{y} = (y_1, y_2, ..., y_n)$, the Euclidean distance between $\vec{x}$ and $\vec{y}$

is given by:

$$D_E(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \,. \tag{7}$$

The Chebyshev distance calculates the maximum absolute distance along any single coordinate dimension:

$$D_C(\vec{x},\vec{y}) = \max_i \left(|x_i - y_i|\right) \,. \tag{8}$$

The Manhattan distance (commonly referred to as the taxi cab distance) is the sum of the absolute distance along every coordinate dimension:

$$D_{Man}(\vec{x},\vec{y}) = \sum_{i=1}^{n} |x_i - y_i| \,. \tag{9}$$

Finally, the Minkowski distance is a generalization of the Euclidean and Manhattan distances:

$$D_{Min}(\vec{x},\vec{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}} \,. \tag{10}$$

Note that the case of $p = 1$ represents the Manhattan distance, $p = 2$ is the Euclidean distance, and $p = \infty$ is the Chebyshev distance.

Less commonly used is the Mahalanobis distance, which accounts for the distribution of the training data. Though harder to interpret and calculate, the Mahalanobis distance can be useful when the variables being measured are highly correlated to each other [74]. The Mahalanobis distance between two points is calculated as:

$$D_{Mah}(\vec{x},\vec{y};Q) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})} \,, \tag{11}$$

where Q is the probability distribution of the training data and S is the positive definite covariance matrix.

Figure 26 shows two examples of a normalized 2-D data distribution with a binary classification. For both, a decision boundary was identified by using isolation forests [75]. Built on decision trees, isolation forests are unsupervised learning algorithms used to identify anomalies. The new data points are indicated by the outermost large blue dots. The Euclidean and Chebyshev distances between each new data point and its closest orange data point (as determined based on the Euclidean distance) are displayed.

Each of the discussed metrics entails its own set of strengths and weaknesses. When the data dimension, n, is high, the Minkowski distances for $p > 1$ are less effective [76]; furthermore, the fractional p values are non-intuitive. The Mahalanobis distance is heavily influenced by all relationships between variables; therefore, it sometimes produces non-intuitive distance results. The Chebyshev distance is both simple and intuitive. In the application area of diagnostics, any one variable being outside the training distribution is enough to cause problems. However, using the Chebyshev distance to identify the nearest training data point may not afford sufficient discrimination, particularly when ordinal variables are included. For the remainder of this report, the Euclidean distance is used to determine the closest point in the training dataset, then the Chebyshev distance is calculated between the two points. This distance is then compared to the confidence of the regression algorithm and to the error in prediction.

4.4.2 Example

To demonstrate the benefits of considering the distance between a new data point and the training dataset, this section presents an example based on the operational data described in Section 3. A regression model

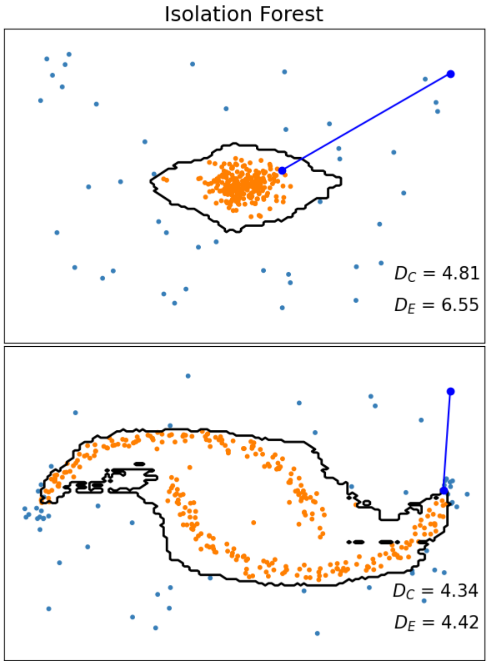


Figure 26. Distance measures applied to two different data distributions.

was built using different ML techniques, then evaluated for a dataset of outliers. The confidence of the regression algorithm, the error in prediction, and the distance from the closest training data point were all compiled for evaluation.

The dataset used for this example includes 3 1/2 years' worth of data associated with an NPP feedwater and condensate system. The specific components being monitored are a condensate pump and its respective motor. The variable being predicted is the motor stator temperature. The predictor variables for the model are: the change in temperature across the water intake and output (i.e., delta temperature (DT)), MIB temperature, Motor outboard bearing (MOB) temperature, motor current, WoY, and pump health status (i.e., healthy or faulted). The week-of-the-year variable was used to capture seasonal effects on the variables, as nominal temperatures tend to vary significantly between summer and winter months. Over the 3 1/2-year period, several planned and unplanned outages occurred, with most of the unplanned ones due to WBF.

With this dataset, we aimed to explore the relationship among model certainty, distance from the closest training data point, and prediction error. This is important because, as average temperatures increase summer after summer [77], the historical data may not adequately reflect the current reality. Furthermore, the emergence of new fault modes can also result in data that lie substantially outside the training dataset, necessitating model extrapolation. The goal of this work is to provide the end user with enough information to confidently determine when the model should be used, and when additional analysis is required.

Because this exercise focuses on predicting outliers, no effort was made to remove them. The data were split into training and testing sets, based on the following threshold applied to the DT variable: any data point with a DT of 15°F or less was placed in the training data, whereas any data point exceeding 15°F was placed in the testing set. This approach was specifically chosen to ensure that every point tested would lay outside the training set. Figure 27 shows the kernel density estimate of the DT for the entire dataset.

After the data split, the training data were standardized by subtracting the mean of the data and then dividing by the standard deviation. The testing data were standardized in like fashion, using a mean and standard deviation that were estimated based on the training data. This was a crucial step for ensuring that

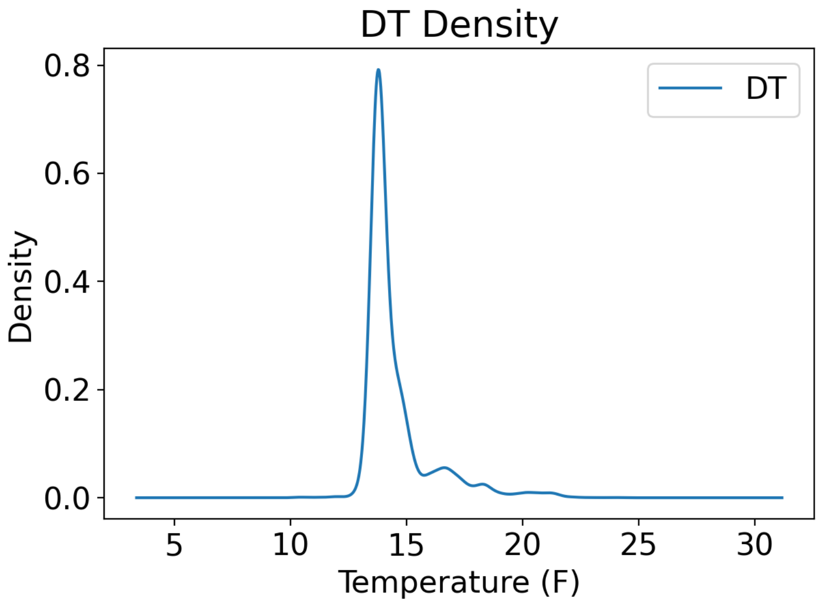


Figure 27. Kernel density estimate of the DT for the entire dataset.

the distance metrics were not dominated by the variable with the largest magnitude. In Figure 28, we see that an isolation forest was able to demonstrate 94% accuracy in predicting which test data points should be considered outliers. Only data points that were relatively close to the training data were misidentified as inliers. Figure 28 also gives the Chebyshev distance for each of the 972 test data points, which were colored by the isolation forest in accordance with their inlier/outlier status—with the x-axis categories indicating which variable was used to determine the Chebyshev distance.

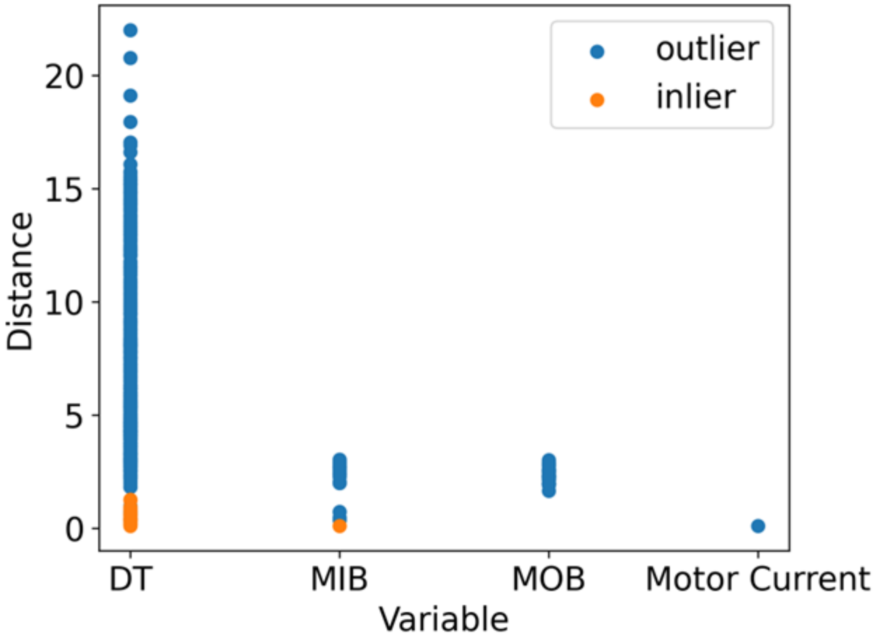


Figure 28. Chebyshev distance of all test data points and the parameter used to calculate it. Each data point was colored by isolation forest in accordance with its inlier/outlier status.

By employing Random Forest (RF) and Support Vector Regression (SVR) algorithms, the training dataset was used to generate regression models. Each model form considered was initially trained on the inlier dataset composed of data points featuring DTs of less than 15°F, then tested on outliers whose DTs exceeded 15°F. A confidence interval was generated for each model to evaluate the relationship among model confidence, prediction error, and distance from the closest training data point.

4.4.3 Random Forest Regression

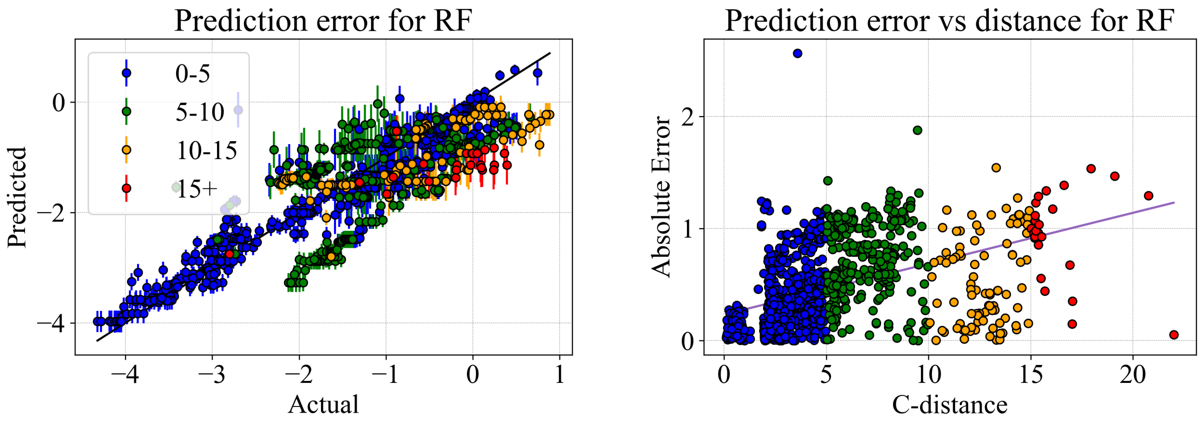
RF is a type of ML algorithm commonly used for classification or regression, and is based on utilizing an ensemble of decision trees [78]. The simplicity of using a single decision tree is offset by its tendency to overfit the data and its vulnerability to variability and noise. RF overcomes these issues by employing bagging aggregation to generate many decision trees. Each individual decision tree is built based on a random sample (with replacement) of the training dataset (i.e., bootstrapping). The RF then aggregates the outputs from the many decision trees and uses the count to produce a classification, or uses the average to produce the final regression value. RF is readily available on many software platforms, including Python's scikit-learn library. RF is commonly chosen because it is accurate, easy to implement, and has built-in feature importance metrics thanks to how the decision trees are structured.

For the example dataset, an RF regression model generated 1,000 decision trees, and the 95% confidence interval for each test data point was generated using Natural Gradient Boosting for Probabilistic Prediction [79]. Figure 29a plots the model prediction against the actual data, along with the 95% confidence interval for the prediction. Recall that each test data point being predicted is an outlier, meaning that model performance is expected to be low. Due to the standardization applied to each variable, the x- and y-axes represent a dimensionless z-score (i.e., how far the actual or predicted motor stator temperature lies from the mean of the training dataset in terms of standard deviation). The different colors indicate the Chebyshev distance between each data point and its nearest training data point. Figure 29b plots the absolute value of the prediction error for normalized motor stator temperature against each data point's Chebyshev distance, along with the 95% confidence interval. Due to the nature of the dataset, which may contain abrupt starts and stops due to maintenance or outages, certain temperature regions are not well represented.

Figure 29b is summarized in Table 11. This table shows the percentage of data points whose actual value lies within the 95% confidence interval for the prediction. The Distance column represents the Cheybyshev distance ranges. In general, the model makes more accurate predictions when the test points are closer to the training dataset, as indicated by the positive slope of the best fit trend line in Figure 29b. This is also demonstrated by the large percentage of close estimates in the 0–5 range in Table 11, whereas there are fewer correct estimates in the 15+ range. The increase in close estimates in the 10–15 range is likely due to the data distribution sampling.

4.4.4 Support Vector Regression

SVR is a type of powerful supervised ML algorithm used for predicting values. In general, SVR uses a kernel function to transform the data to a higher dimension, then generates a decision boundary (i.e., hyperplane) that is used to predict a continuous parameter. The closest data points on either side of the hyperplane are the support vectors. Most multivariate regression models simply generate a hyperplane that minimizes the error between the real and the predicted value, but SVR generates a hyperplane that minimizes the error between the real and the predicted value only for the subset of data that exceed the threshold distance (ϵ) from the hyperplane. The kernel function transform makes hyperplane optimization more efficient. The kernel functions can be selected to accommodate nonlinear relationships. Common kernel functions include



(a)

(b)

Figure 29. RF results: (a) predicted versus actual values (after standardization) for a set of outliers; (b) prediction error versus Chebyshev distance from the nearest training data point (after standardization) for a set of outliers.

linear, radial basis, and polynomial functions.

To gauge the certainty of the SVR model, a modified bagging approach was utilized. A hundred models were generated based on a random selection from three regularization parameters ($C \in [0.5, 1, 2]$), three epsilon values ($\epsilon \in [0.05, 0.1, 0.2]$)), and a random sample (with replacement) of training data that is equal in size to one-third of the entire set of training data. For each test data point, the average of these 100 models was taken to be the predicted value, and the standard deviation was used to calculate the 95% confidence interval. Examples using each of the aforementioned types of kernel functions are given here.

The linear kernel is simply the dot product of the two points. It can be used when a linear model adequately represents the response variable:

$$K_{Linear}(\vec{x}, \vec{y}) = \vec{x}^{\mathsf{T}} \vec{y}. \tag{12}$$

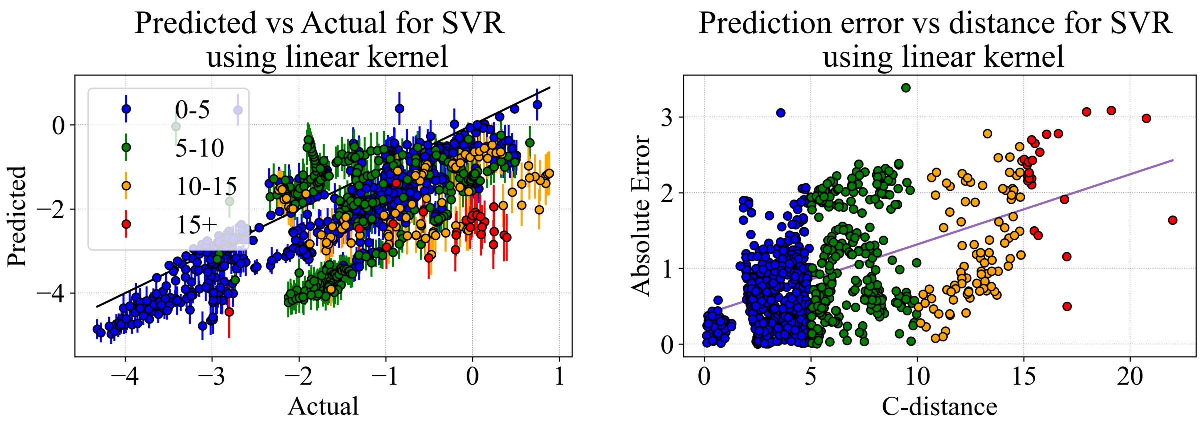
SVR with a linear kernel was used to generate a model and the corresponding 95% confidence interval for the predictions. For the test dataset, Figure 30a shows the predicted versus the actual data points, along with the confidence interval of one standard deviation. The color of each data point indicates its Chebyshev distance from the closest training data point. Figure 30b plots the absolute value of the prediction error, along with the confidence interval, against each data point's Chebyshev distance. Figure 30b is summarized in Table 11. It is interesting to note that in this example the linear SVR generally overpredicts motor stator temperature for the test data set. Overall, the linear SVR model behaves similarly to the RF model, with slightly fewer test data points lying within the 95% confidence bounds at all Chebyshev distance ranges.

For degree d, the polynomial kernel function is defined as:

$$K_{poly}(\vec{x}, \vec{y}) = (\vec{x}^{\mathsf{T}} \vec{y} + c)^d , \tag{13}$$

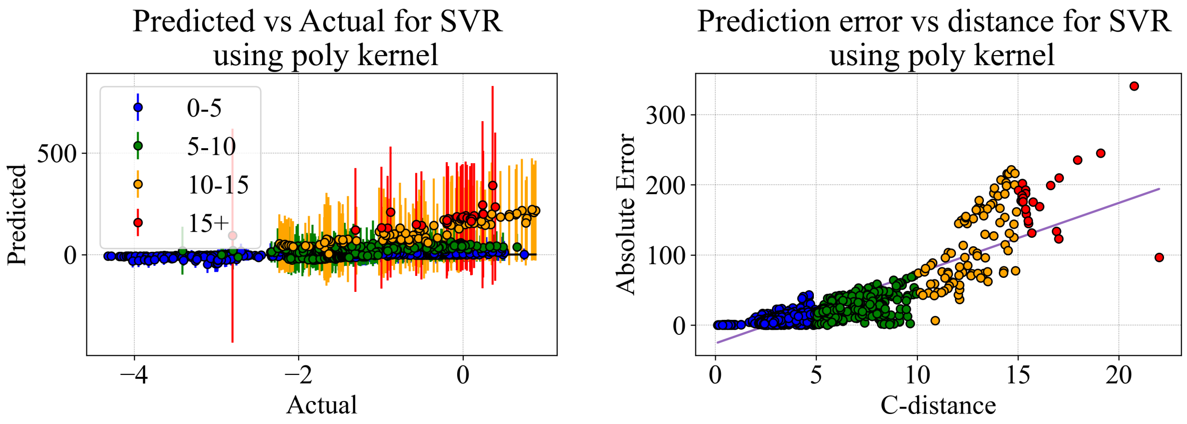
where $c \geq 0$ is a kernel parameter. When $c = 0$, the kernel is homogeneous.

The results of the homogeneous SVR polynomial predictions, as shown in Figure 31 and Table 11, suggest that the polynomial kernel does not handle outliers well. The confidence intervals of the furthest data points are too broad to be of any use. This is not entirely surprising, as even with only one predictor,



(a) (b)

Figure 30. Linear SVR results: (a) predicted versus actual values (after standardization) for a set of outliers; (b) absolute prediction error versus Chebyshev distance from the nearest training data point (after standardization) for a set of outliers.



(a) (b)

Figure 31. Polynomial SVR results: (a) predicted versus actual values (after standardization) for a set of outliers; (b) absolute prediction error versus Chebyshev distance from the nearest training data point (after standardization) for a set of outliers.

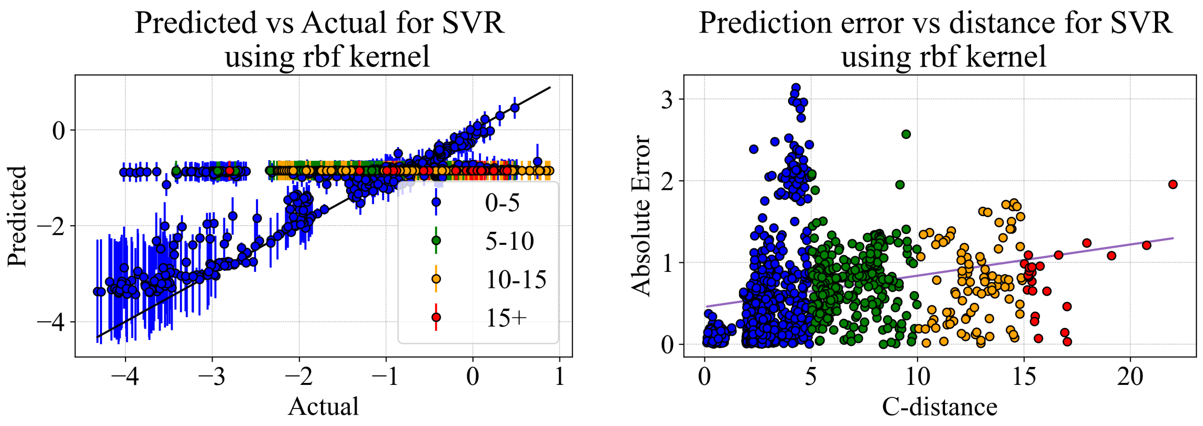
a polynomial that fits beautifully within the training data may behave poorly for extrapolation, due to the presence of local minimum or maximum points.

The radial basis function kernel is one of the most widely used kernels, thanks to its resemblance to the Gaussian distribution. It is defined as:

$$K_{RBF}(\vec{x},\vec{y}) = \exp\left(\frac{-D_E^2(\vec{x},\vec{y})}{2l^2}\right) , \tag{14}$$

where the parameter l represents the length scale for the kernel. The radial basis kernel function is an infinite-dimensional version of the polynomial kernel.

The results generated using the radial basis kernel, as shown in Figure 32 and Table 11, reveal some



(a)

(b)

Figure 32. RBF SVR results: (a) predicted versus actual values (after standardization) for a set of outliers; (b) absolute prediction error versus Chebyshev distance from the nearest training data point (after standardization) for a set of outliers.

interesting behavior. The model seems to return a default value for the data points farthest from the training data, resulting in a horizontal lineup of data points, along with relatively small 95% confidence bounds (see Figure 32a). This is a case in which the distance information may provide end users with valuable feedback regarding the fact that the model results may be inaccurate due to extrapolation.

Table 11. Percent of the actual values that are within the 95% confidence interval of the prediction.

Distance	RF	SVR linear	SVR polynomial	SVR RBF
0–5	0.429	0.352	0.525	0.178
5–10	0.105	0.047	0.989	0.142
10–15	0.250	0.120	1.000	0.076
15+	0.040	0.120	1.000	0.000

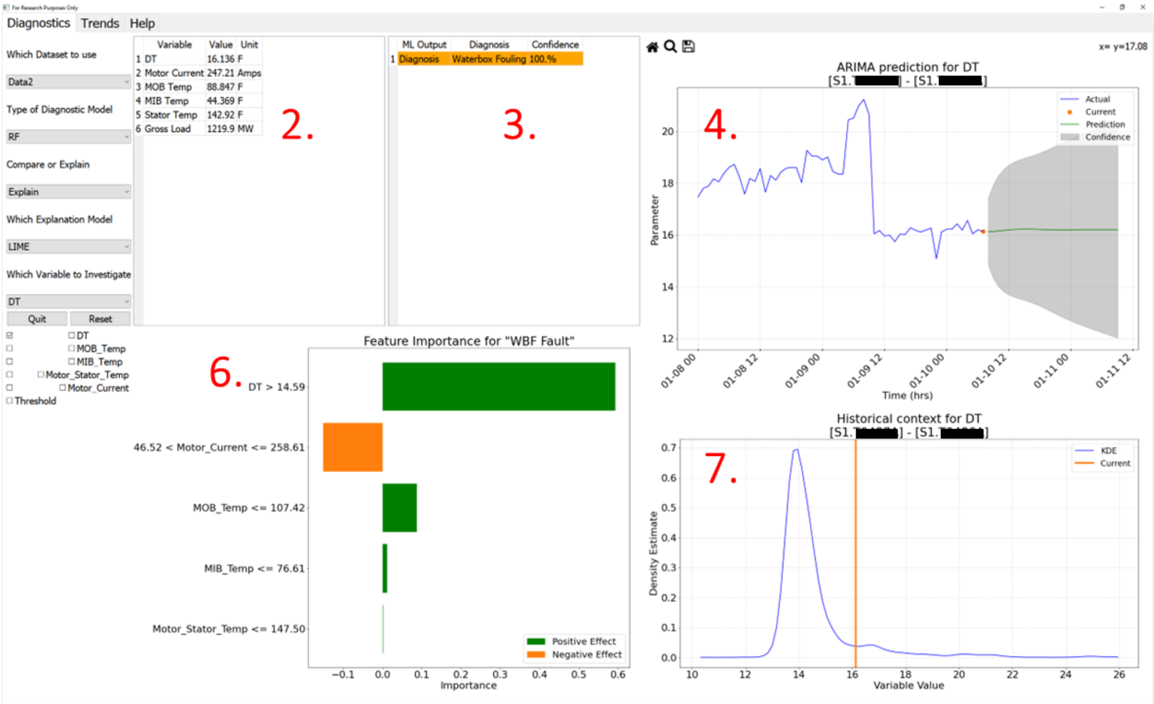
The purpose of this example was to demonstrate that quantifying the distance between a data point of interest and the closest data point in the training dataset equips end users with useful contextual information. When the considered data point was sufficiently far from the nearest training data point, all models performed poorly, and most were overconfident in their results. Thus, merely knowing the confidence of the model is not sufficient evidence to support making a decision based on model results. This issue is particularly important in high consequence industries such as the nuclear industry, where explainability and trust are critical to enable broad adoption of AI/ML technologies.

Taken together, this chapter outlined how employing the metric of data novelty should improve model trustworthiness by declaring if the current data are consistent with or close to the training data. Caution must be taken when extrapolating with any model as poor or unexpected performance may occur. By incorporating data novelty into the ML output presentation, a guardrail has been added to notify the user that although the model confidence may be high, it does not have enough evidence within the training data to back up that conclusion. An example of how to present machine-learning outcomes is given in the following chapter.

5 USER-CENTRIC VISUALIZATION

One of the goals for this project was to develop a user-friendly application to put ML capabilities into the hands of maintenance and diagnostics (M&D) operators. Although these operators are knowledgeable in plant parameters and processes, they may not be as familiar with ML techniques and how these algorithms arrive at a certain diagnosis based on a set of features. M&D operators will not blindly follow a ML model's recommendation. To build trust with the M&D operator, the model should be as transparent and explainable as possible. This section of the report covers the development of the M&D app as well as the design choices that were made to further increase usability.

The application consists of three tabs which group relevant features together. The Diagnostics tabs, shown in Figure 33, contains the most features. All of these features serve to aid the operator in understanding what the ML diagnosis was and evidence that supports that conclusion. The Trends tab, shown in Figure 34, shows each of the signals being monitored and provides an estimate of that feature into the near future with 95% confidence. This prediction horizon (number of hours into the future) can be selected by the user. The final tab is the Help tab which contains a help menu that explains every acronym, feature, and model.



Which Dataset to use: Data2

Type of Diagnostic Model: RF

Compare or Explain: Explain

Which Explanation Model: LIME

Which Variable to Investigate: DT

☑ □DT
□ □MOB_Temp
□ □MIB_Temp
□ □Motor_Stator_Temp
□ □Motor_Current
□Threshold

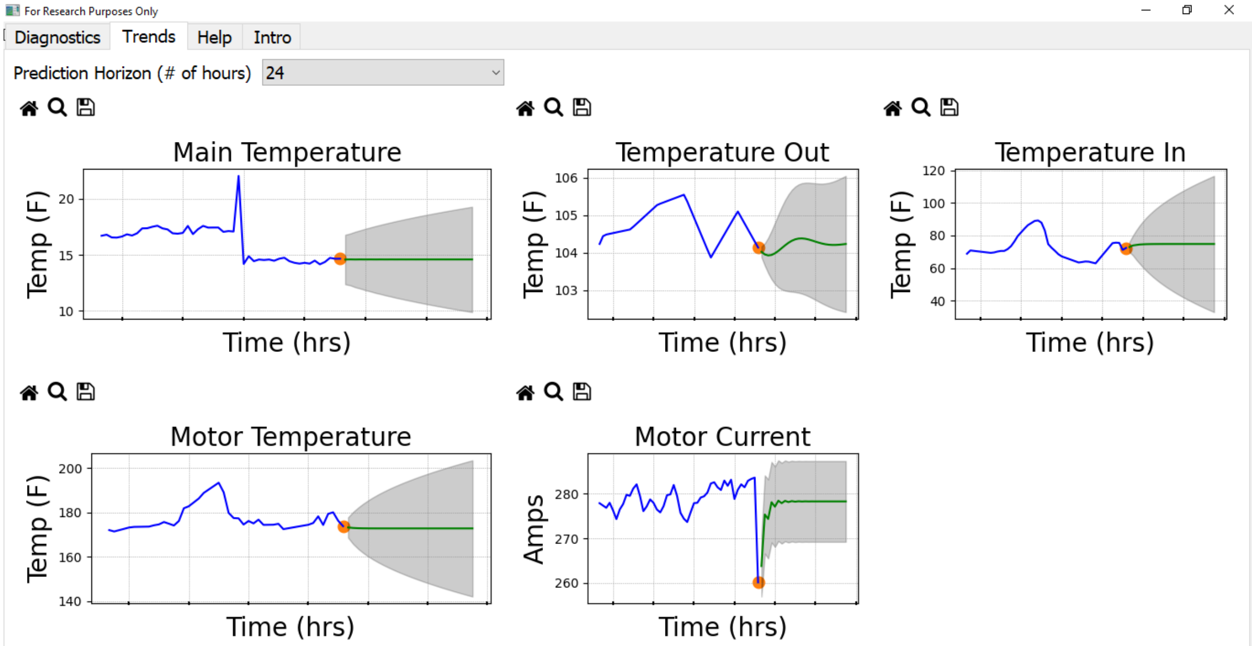
	Variable	Value	Unit
1	DT	16.136	F
2	Motor Current	247.21	Amps
3	MOB Temp	88.847	F
4	MIB Temp	44.369	F
5	Stator Temp	142.92	F
6	Gross Load	1219.9	MW

	ML Output	Diagnosis	Confidence
1	Diagnosis	Waterbox Fouling	100.%

Figure 33. Front page of the M&D operator-version of the application.

The diagnostic tab contains seven areas:

1. User Selection
2. Current Plant Parameters
3. ML Outcomes
4. Current Feature Trending
5. Variables to Compare



Prediction Horizon (# of hours) 24

Figure 34. Trends tab showing the prediction of each of the measured features.

6. Feature Importance and Comparison
7. Historical Context.

The area 1, labeled as the “User Selection,” allows the operator to choose the models, variables, and plots they would like to see in a singular location. Currently, ‘Which dataset to use’ allows the user to select between different test cases to see how the ML models respond to different scenarios. In real use cases, this feature could be slightly altered to allow the user to examine other pumps or components of interest. ‘Type of Diagnostic Model’ currently allows the user to select between using RF and support vector machine for diagnosing potential faults in the system. Additional diagnostic models can be added easily as each of the subsequent explanation models are post-hoc methods which means they are not integrated into the model itself, but rather used on a pre-trained model. ‘Compare or Explain’ allows the user to compare multiple features, as seen in Figure 35, or explain how the model reached its diagnosis using either Local Interpretable Model-agnostic Explanations (LIME) or Shapley Additive Explanations (SHAP). Explanation on the underpinnings of LIME and SHAP can be found at [80] and [81], respectively. ‘Which Variable to Investigate’ allows the user to focus on one variable more intently and this changes areas 4 and 7 in Figure 33.

Area 2 contains the current plant parameters. This allows the operators to assess all features in one place, which can allow for an at-a-glance check. This also aids in the trust-but-verify aspect of the app as the operator may be more familiar with the plant parameters than the ML.

Area 3 contains all ML outcomes. This includes the diagnosis (healthy, WBF, diffuser faults, etc.), the ML confidence in this decision, as well as novelty detection (discussed more in Section 4.3). Extra space is given in this section to accommodate for more ML outcomes such as estimated-time-to-failure or other ML recommendations.

Area 4 contains a figure of the current variable of interest. The title lists the model used to make the 24-hour ahead prediction, Autoregressive Integrated Moving Average (ARIMA), the variable’s name, and the variable’s location inside the data historian. The blue line represents the recorded data that will be

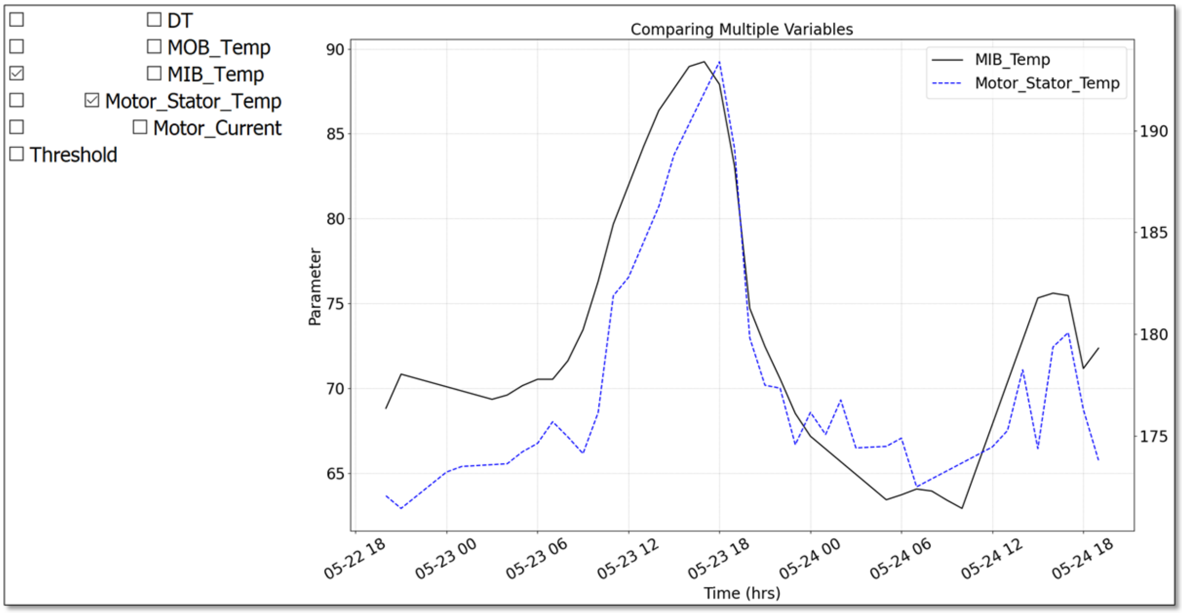


Figure 35. This figure is included in the M&D operator-version of the application and allows the operator to compare multi-signals on the same plot. This is useful to see how or if features are trending similarly.

updated in either a continuous or batch-like manner while the red section represents the 95% confidence of the predicted variable as it is calculated for the upcoming 24-hour period.

Area 5 allows the user to compare multiple features on the same plot. Each variable is given a left and right checkbox. By checking the left box, the variable is plotted and the left axis is scaled accordingly. By checking the right box, the variable is plotted and the right axis is scaled accordingly. This allows for variables of differing magnitudes to be compared side-by-side to quickly see how or if they are trending together. In Figure 35, the MIB temperature and motor stator temperature are both plotted. Although their temperatures are roughly 100 degrees separated, it can easily be seen that they are trending similarly. This section also contains a Threshold checkbox. This plots the failure thresholds, as determined by guidance, for each of the selected variables. This helps to put into perspective whether the component is nearing failure based on a single variable.

Area 6 contains the Feature Importance and Comparison figures. By default, the multi-feature comparison figure is shown, as requested by an M&D operator. When a user wants to learn why the model is behaving the way it is, they can select ’Compare’ to show the post-hoc feature importance metrics of either LIME or SHAP. This figure then list the most important variables at the top, based on how important that variable was to the ML model’s diagnosis. In Figure 33, the most important variable for the WBF diagnosis was DT. With LIME, the figure also explains why is it is contributing to that diagnosis based on certain ranges. DT is seen to be abnormally high by being greater than 14.59, while the motor current is within an acceptable range between 46.52 and 258.61 amps.

Area 7 shows the historical context of the selected feature. This context is shown via a kernel density estimate which shows the distribution of the data used to train the model. In Figure 33, the current DT is sitting on the right side of the distribution so it can be seen that this temperature is relatively high in its historical context.

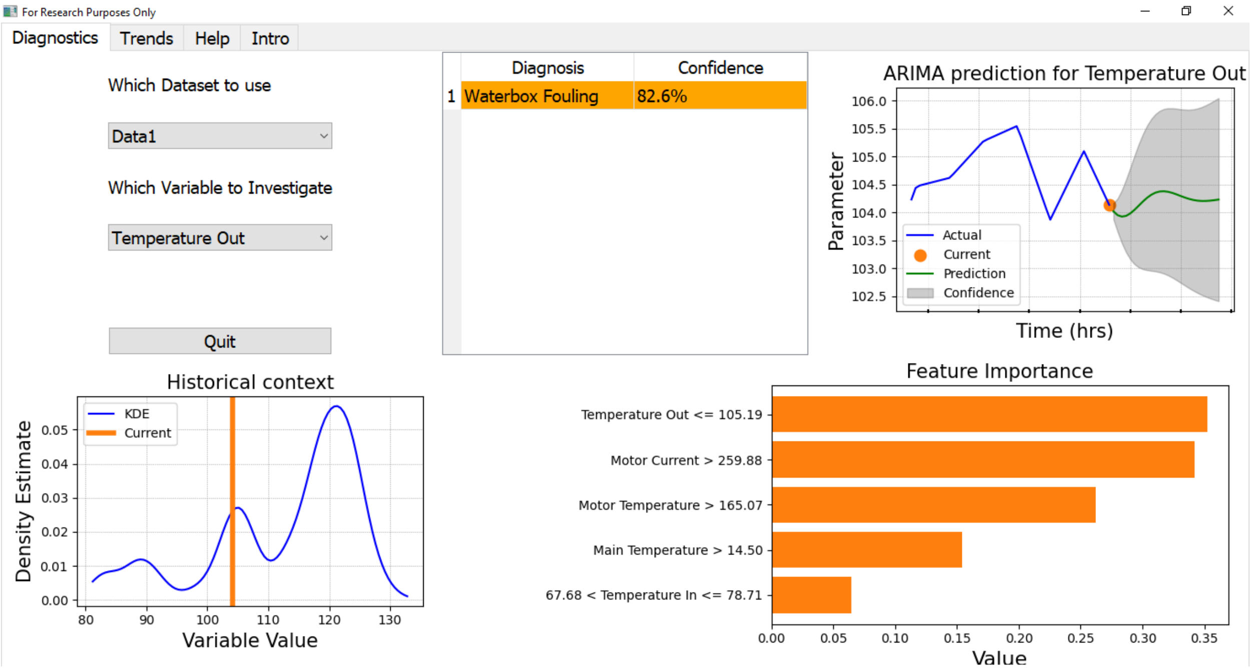
The overall goal of this application is to give M&D operators an additional tool, not to replace the operator. By including things such as explainability metrics, historical context, model confidences, and predictions into the near future, this tool should be able to build trust with the operator so that they would feel comfortable making recommendations based on its conclusions. If the operator did not trust the model's conclusion, there are various ways that the operator could perform an independent analysis through the either the Trends tab, historical context, or the multi-feature comparison tool.

6 TRUSTWORTHINESS

This section covers human interactions with the user-centric visualization to determine if the explainability metrics and app presentation was adequate. This section also describes the trust, but verify approach and how that approach can aid with AI adoption. It ends with a proposed study to measure this approach in future research.

6.1 Human Factors Evaluations

To elicit feedback from a wider audience to improve the app further, a reduced version of the app, shown in Figure 36, was created and shown at the Nuclear Plant Instrumentation Control & Human-Machine Interface Technologies (NPIC&HMIT) 2023 conference. Since this new audience has a different set of skills to the intended M&D operators and would be less familiar with the plant processes and parameters, the NPIC-version of the app focused on the explainability aspects while reducing the amount of plant-specific features and overall usability. This more focused app was shown, explained, and used as a reference for the survey shown in Figure 37. Questions 2–5 of the survey were used to solicit feedback to test the main hypothesis: the app contained sufficient explainability that the users would trust the algorithm.

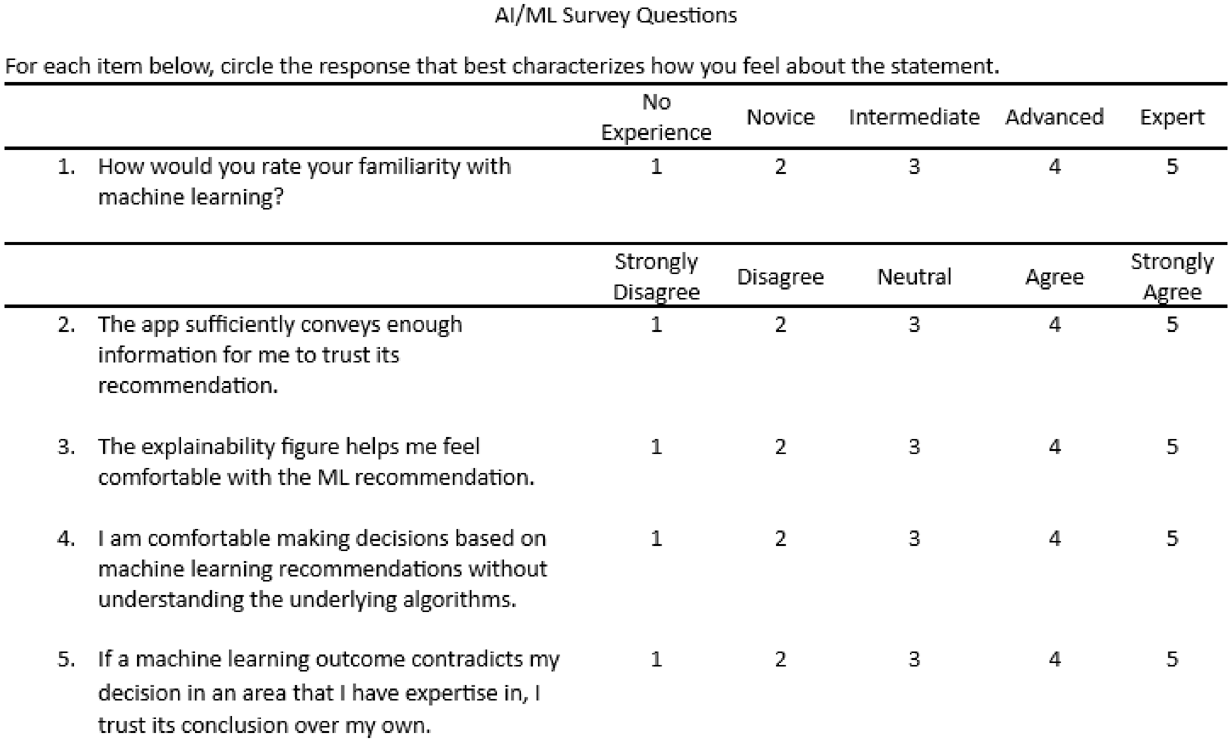


Which Dataset to use: Data1

Which Variable to Investigate: Temperature Out

	Diagnosis	Confidence
1	Waterbox Fouling	82.6%

Figure 36. Front page of the NPIC application. This app has been tailored and focused for a diverse conference audience.



AI/ML Survey Questions

For each item below, circle the response that best characterizes how you feel about the statement.

	No Experience	Novice	Intermediate	Advanced	Expert
1. How would you rate your familiarity with machine learning?	1	2	3	4	5

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
2. The app sufficiently conveys enough information for me to trust its recommendation.	1	2	3	4	5
3. The explainability figure helps me feel comfortable with the ML recommendation.	1	2	3	4	5
4. I am comfortable making decisions based on machine learning recommendations without understanding the underlying algorithms.	1	2	3	4	5
5. If a machine learning outcome contradicts my decision in an area that I have expertise in, I trust its conclusion over my own.	1	2	3	4	5

Figure 37. Survey given at the 2023 NPIC conference to gain feedback on the app development.

Given the diverse audience expected at NPIC, question 1 was used to determine how much experience each participant had with ML. Figure 38 shows the breakdown of the 32 survey responses received at NPIC by level of expertise with ML. Only two participants indicated they had no experience with ML. A total of 19 respondents fit into the novice and intermediate groups, while 11 fit into the advanced and expert groups. The team hypothesized that there would be an interaction between familiarity and trust in that the level of familiarity with ML would influence the level of trust in the ML recommendations. However, this was not the case. The distributions of the responses based on level of expertise were similar to each other, indicating level of familiarity did not have a strong influence on whether the user would trust the recommendations based on the information displayed.

Question two explored the hypothesis that the application displayed enough information (i.e., machine learning confidence, feature trending, historical context, and feature importance) for the user to trust the model's recommendation. The responses seen in Figure 39 show that the hypothesis was supported. Five total participants expressed a neutral opinion, 23 agreed, and four strongly agreed. The distributions of the responses based on level of expertise are similar to each other, indicating level of expertise did not have a strong influence on whether the user would trust the recommendations based on the information displayed.

Question three focused on ascertaining if the explainability figure (labeled as the feature importance panel in Figure 36) helped the user feel comfortable with the ML recommendation. The responses seen in Figure 40 show that none of the participants disagreed, four total participants had a neutral opinion, 21 agreed, and seven strongly agreed that the explainability figure increased user trust.

Question four asked participants to indicate how comfortable they would be making decisions based on ML algorithms without fully understanding the underlying algorithms. The responses seen in Figure 41 show that none of the participants strongly disagreed, nine disagreed, 12 were neutral, nine agreed, and only two strongly agreed. Interestingly, the two participants who strongly agreed were both experts in ML.

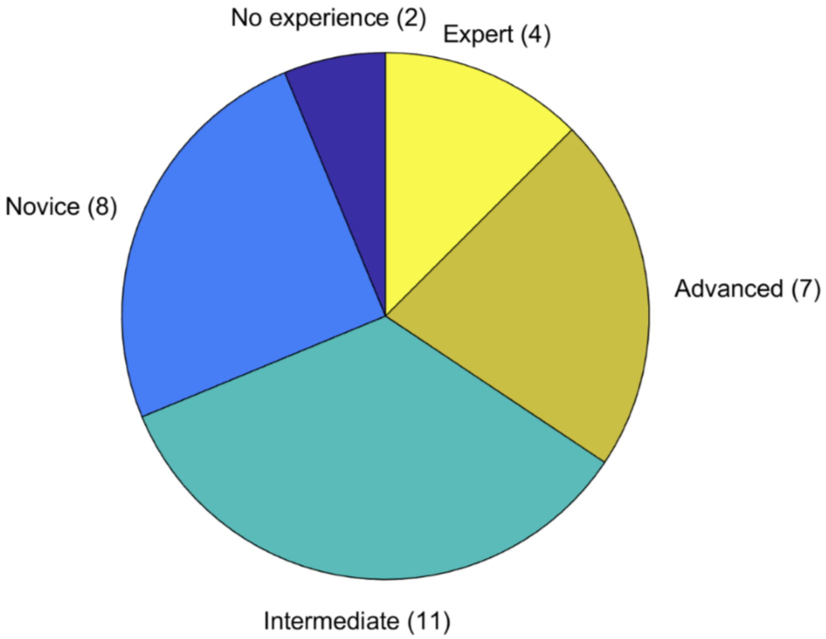


Figure 38. Survey respondents' ML familiarity.

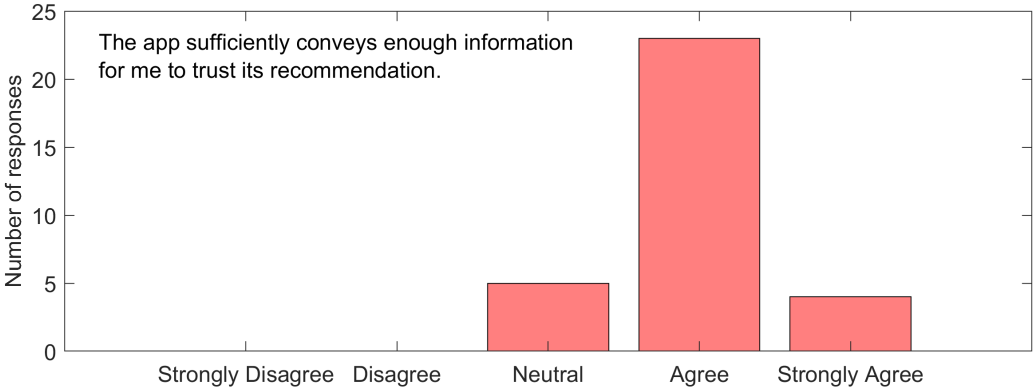


Figure 39. Responses to Question two on whether the app conveyed enough information.

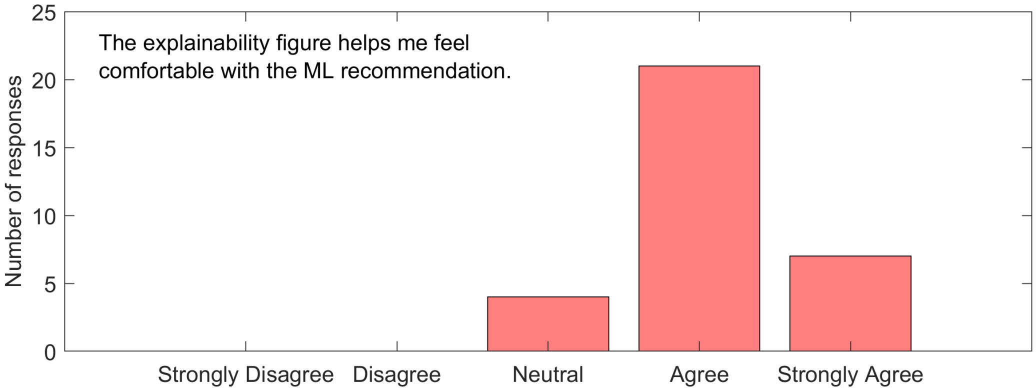


Figure 40. Responses to Question three on whether the app conveyed enough information.

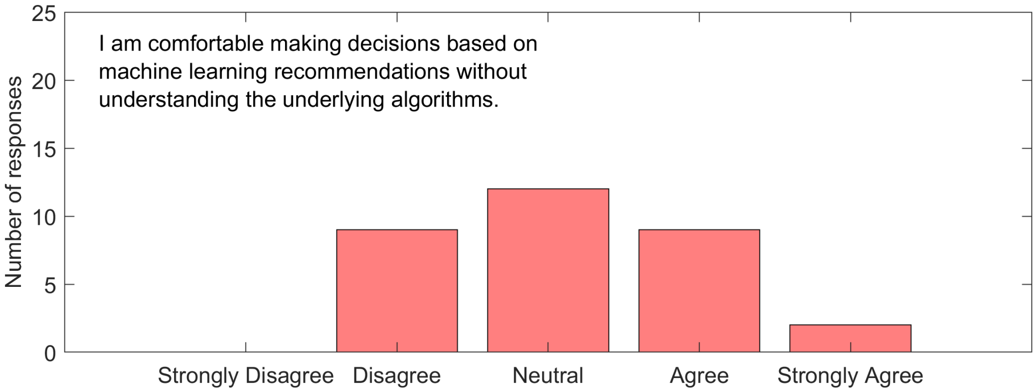


Figure 41. Responses to Question four on whether participants are comfortable without knowing how the underlying algorithm works.

The last question was focused on whether the participants would trust the ML outcome when it contradicts with their own conclusions. The responses seen in Figure 42 show that two participants strongly disagreed, seven participants disagreed, 16 were neutral, five agreed, and two strongly agreed that they would trust the ML algorithm over their own conclusions. The distributions of the responses by level of expertise show that the advanced and expert users had the largest range of responses, and their responses were fairly evenly split amongst the five categories. Additionally, one respondent indicated that the first time such a discrepancy occurred, they would dig into the data more deeply to verify the ML output was indeed correct. They would then be more willing in the future to accept the ML output.

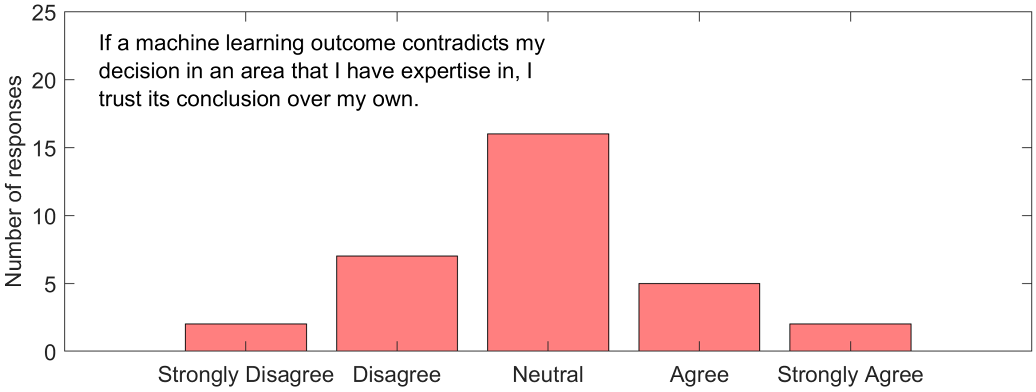


Figure 42. Responses to Question five on whether participants would trust the ML recommendations over their own if it contradicted them.

A correlation analysis was performed on the answers to questions 2–5. Moderately positive correlations of 0.6034 and 0.572 were seen between the responses to questions 2 and 4, and to questions 4 and 5, respectively. Contextually, this is perhaps common sense. If the application supplies sufficient information to make the user trust its recommendation, then they would likely be comfortable making decisions, even without understanding the details of the algorithm. Additionally, if the user is comfortable making decisions based on the ML algorithm, even without understanding the details of the algorithm, then they might be more likely to trust an ML conclusion that conflicts with their own.

However, this survey had a limited number of respondents and likely did not contain the app's targeted final user (i.e., M&D operators). In the future, more responses from the target user group will be obtained through additional surveys and testing. Overall, the conference responses indicate a positive response to the app and trust in the ML technology.

6.2 Trustworthiness for Artificial Intelligence

Reference [82] presents three levels of trust from human level to AI level. Each level is briefly discussed here.

1. Human level: In general, trust is perceived as a relationship between two or more entities bounded by certain expected principles, facts, or contract. In this level, one entity is trustor and another entity is trustee. The relationship between trustor and trustee is established and maintained as certain contracts outlining terms and conditions based on association between them.

2. Computer level: In computers, trust applies to hardware and software systems (trustee) as well as their interaction with humans (trustor) and the physical world. It is understood that both hardware and software systems are subject to different threats. Therefore, trust at computer level is governed by certain properties that include reliability, safety, security, privacy, availability, and usability. Some literature also suggests trust at the computer level can be established via a formal verification process [83].

3. AI level: Similar to human and computer trusts, the trust in AI (trustee) is based on beliefs or perceptions of its trustworthiness by the user (trustor) which in its current form is function of how it is perceived by the user in terms of technical trustworthiness characteristics [84]. For instance, the Organization for Economic Cooperation and Development promotes five principles to ensure trustworthy AI:

1. Inclusive growth, sustainable development, and well-being
2. Human-centered values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability.

Additional principles have been proposed such as trustworthiness [83], acceptance [85], predictability, and performance [86].

By identifying properties or measures of trustworthiness, there is a potential to benchmark how well AI/ML solutions compare to one another. Tidjon and Khomh [82] reviewed a significant number of articles associated with trust in the context of AI-based systems to understand what it means for an AI system to be trustworthy and to identify actions that need to be undertaken to ensure AI systems are trustworthy. They identified 12 trustworthy properties: transparency, privacy, fairness, security, safety, responsibility, accountability, explainability, well-being, human rights, inclusiveness, and sustainability that are widely adopted in establishing trustworthiness of an AI-based system. For details on each property, see [82].

For the nuclear industry to adopt AI-based systems, trustworthy principles across all the three levels —from human to AI—needs to be established. The following ”trust but verify” study focuses on the AI level of trust.

6.3 Trust but Verify Approach

A previous effort [10] followed a human factors engineering design and evaluation process for a single iteration of the explainable AI/ML interface from which the following lessons learned were detailed:

- Although integrating automated technologies within process control industries has shown a reduction in human error and an improvement in labor-intensive work systems [87], an essential but occasionally overlooked factor that directly impacts the success of automation integration is the ability and willingness of NPP personnel to rely on these technologies. Therefore, the absence of a framework for how to increase human reliance on automation will directly limit the benefits of incorporating automation in the first place.
- Future research could embrace the influence of nuclear safety culture by developing a study wherein AI/ML models are used as an additional source of data that has the potential to increase decision making capabilities.
- There is an abundance of opportunity to incorporate automated technologies such as AI/ML models into the nuclear industry but identifying how these models are interpreted and relied on by humans is essential to realize the full intended benefits of automation integration. Further research is needed to fully understand this concept, and lessons learned from this work as well as influences of nuclear safety culture should be reflected.

The "trust, but verify" concept is derived from an observation of nuclear safety culture (i.e., NPP personnel do not rely on a singular source of data to make a decision) and how that might limit the effectiveness of integrating new technologies such as ML models. The purpose of integrating ML models is to automate tedious and labor-intensive work to simplify and streamline NPP personnel task load. However, forcing NPP personnel to adopt a new decision-making model that is contradictory to the general safety culture will likely lead to a rejection of the models themselves. Therefore, accepting the idea that NPP personnel rely on the ability to evaluate multiple sources of information to make a decision provides the opportunity to leverage their existing decision-making model without limiting the effectiveness of ML model integration.

This concept was also demonstrated within the preliminary interface evaluation [10]. One of the key takeaways was that participants used the WBF model as an additional source of decision-making data as opposed to a singular source. In other words, they were more likely to "trust" the model once they had the opportunity to "verify" additional information included in the interface. Therefore, researchers set out to develop a study that more closely aligns with the concept of "trust but verify."

For the purposes of this study, "trust but verify" is defined as a participant's process to make a decision. This includes participants exhibiting both trust (i.e., adhering to the model's recommendation) and verify (i.e., evaluate additional sources of information) behaviors to be considered.

6.3.1 Future Study: "Do ML Models Improve Decision-Making Capabilities?"

For the purposes of this proposed study, "do ML models improve decision-making capabilities" is defined as a participant's process to make a decision. This study differs from the preliminary evaluation concerning the interface design feature for the model recommendation in an important way that is relevant for measuring the decision-making process. Instead of including a confidence interval (e.g., low, medium, or high) combined with a "waterbox" or "healthy" status, the model recommendation is simplified to include an action of “delay maintenance” or “perform maintenance.” This change is based on the lesson learned from the previous study of further simplifying what action the model is recommending. This way, any potential confusion concerning participants misinterpreting the model’s recommendation (i.e., confidence interval combined with status) is eliminated.

The proposed structure for this study is a within-subjects comparative analysis (Figure 43). The independent variable is the presence of the ML model recommendation which includes two categories: an interface with the model recommendation present and an interface without the model recommendation (see Figure 44). Otherwise, the interfaces and the backend ML are identical. All participants will be presented with both Condition A (interface with model) scenarios and Condition B (interface without model) scenarios with a combined total of 12 scenarios, and then asked to make a decision (i.e., to perform or delay maintenance). The order of treatment presentation will be counterbalanced to control for any learning effects.

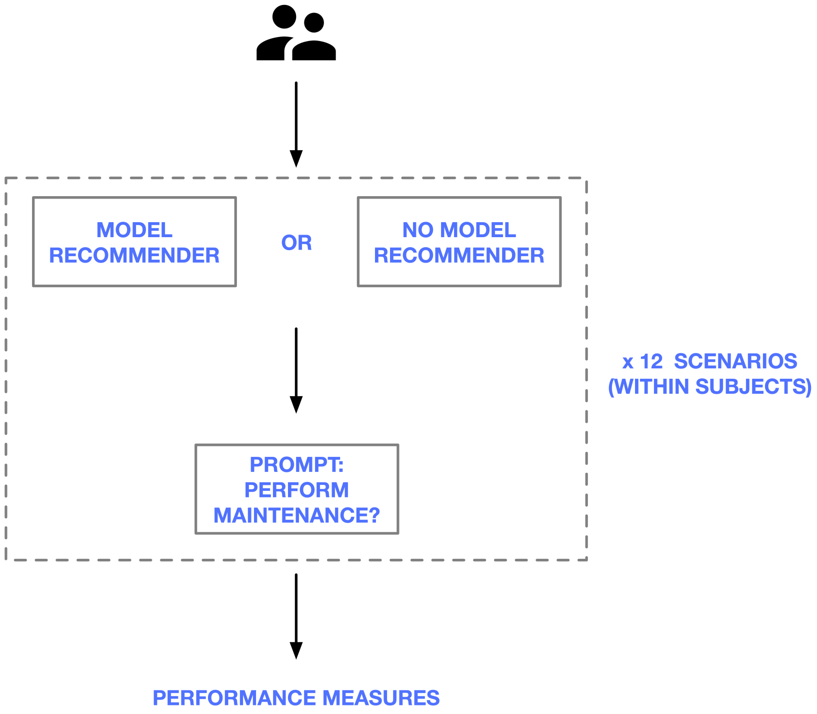
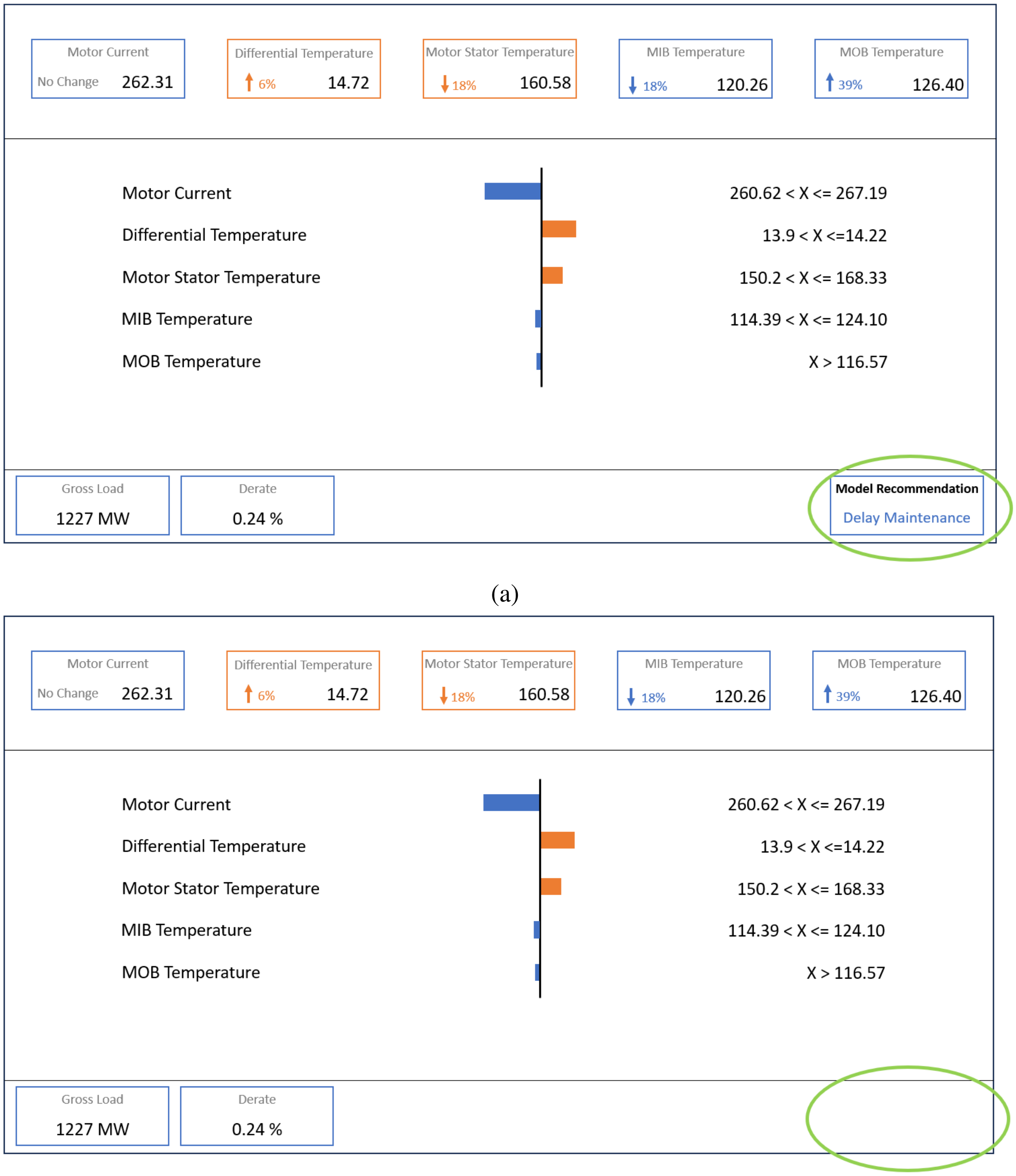

MODEL RECOMMENDER
OR
NO MODEL RECOMMENDER
x 12 SCENARIOS (WITHIN SUBJECTS)
PROMPT: PERFORM MAINTENANCE?
PERFORMANCE MEASURES

Figure 43. Proposed study structure.

Each scenario across both conditions will include a WBF maintenance prompt: to perform maintenance or to delay it, with the independent variable being the inclusion or exclusion of the recommendation. The WBF scenarios and associated parameters will be identified with the help of subject matter experts/utility stakeholders. The goal is to provide decision scenarios that participants would be able to discern on their own (i.e., to perform or delay maintenance) however, participants may take a lot of time or wrongly recommend/delay maintenance. The primary purpose of integrating AI/ML is to improve speed and decision reliability. Therefore, WBF parameters should not be straightforward enough that participants will not solely rely on the model to make a decision (i.e., they will still verify additional information) but will instead seek additional decision-making data, hence “do ML models improve decision-making capabilities.”

In Figure 44, examples of both interface conditions are shown. On the left side is condition A (interface WITH model recommendation) and on the right side is condition B (interface WITHOUT model recommendation. The green oval on the bottom right of each interface highlights the location of the interface that displays the model recommendation location (note that condition B does not include a model recommendation as part of the study structure).



(a)

(b)

Figure 44. Interface with (a) model recommendation and (b) without model recommendation.

The primary dependent variable is the level of confidence participants have that their decision is correct (i.e., decision to delay or perform maintenance). If participants' confidence is higher when completing scenarios with the model recommendation compared to scenarios without the model recommendation, it can be reasonably interpreted that ML models improved a participant's decision-making capabilities. Following each scenario, participants will be asked to rate their confidence on a sliding scale of one (i.e., not at all confident) to ten (i.e., completely confident). Additionally, experimental data concerning scenario duration (i.e., speed) of participants will be captured. Performance metrics concerning accuracy (i.e., whether or not the participant action to delay or perform maintenance was correct) will also be collected. Each metric will accumulate to an overall decision-making score. The main hypotheses to be tested are listed as follows:

- Participants will record higher confidence in their decision-making capabilities during “with model” scenarios compared to “without model” scenarios.
- Participants will record faster speed in their decision-making capabilities during “with model” scenarios compared to “without model” scenarios.
- Participants will record better accuracy in their decision-making capabilities during “with model” scenarios compared to “without model” scenarios.

These hypotheses are based on the researchers’ observation that although ML models will not be used as a singular source of decision- making data, they will be used as an additional source of decision-making data which will heed stronger confidence in their overall decision accuracy.

If the listed hypotheses are confirmed, it will demonstrate that ML models increase the confidence level of NPP personnel decisions. It will also validate the observation that ML models should be integrated in a way that reinforces the current decision-making process of NPP personnel.

This study could also pave the way for future research that evaluates how to best display model recommendations to NPP personnel. For example, NPP personnel working the same job have varying degrees of expertise conducting the manual processes that ML models would automate. An additional hypothesis would be that less familiar (i.e., newer to this type of work) participants will display the highest level of confidence (compared to other participants) within “with model” scenarios. It is assumed that less experienced participants will be more likely to rely on ML models to make a decision and would benefit from design features that support this level of reliance such as model recommendations and varying degrees of explainability and transparency. Future studies can evaluate this hypothesis by collecting demographic data concerning years of experience.

7 SUMMARY AND PATH FORWARD

This report examined the potential of AI and ML technologies in the nuclear industry to enhance decision-making, reduce O&M costs, and increase the efficiency of NPPs. The report identifies various barriers to the adoption of AI and ML in the industry, including technical, economic, readiness, and stakeholder challenges. To address these barriers, the report proposes solutions focused on improving the explainability of ML models to build trust among end users.

The trade-off between ML performance (accuracy) and explainability is discussed, with highly accurate methods like deep learning being less explainable, while more explainable methods like linear ridge classifier may sacrifice some accuracy. The report emphasizes the importance of data novelty and the value of new information in assessing explainability and trustworthiness. Novelty detection helps determine how new data aligns with the existing training data, while the value of new information could contribute to recommendation systems as additional collected information would strengthen the confidence of machine learning outcomes.

This report described the development of a copyrighted user-centric visualization aligned with the human-in-the-loop approach. This visualization presents different levels of information and can be tailored to individual user credentials to instill user confidence in ML methods. It includes explainability metrics for the presented ML models. Feedback from 32 users with varying ML expertise supports the hypothesis that the app’s explainability fosters user trust in the algorithm.

The trust, but verify framework was proposed as a potential approach to establish user trust in AI. Inspired by nuclear safety culture, which relies on multiple data sources for decision-making, this framework involved building trust from the human level to the AI level. The user-centric visualization plays a critical role in achieving both explainability and trustworthiness of AI.

However, there are still other barriers that need to be addressed for adoption of AI/ML in nuclear, namely regulatory and stakeholder readiness. As new and innovative solutions are created to solve technical challenges, these solutions need to be introduced with the NRC's approval and stakeholder usage in mind. The NRC does have a 5-year strategic plan for AI which outlines the agency's preparations for readiness to review licensee submissions that employ AI technologies. It's recommended to engage with the NRC early and often. The developed solutions must also consider the human-in-the-loop and consider their technical background or provide ways of making the solutions more explainable. The stakeholders will also need to hire or train staff as technology continues to improve and become more incorporated into everyday plant tasks.

REFERENCES

[1] S. W. Foon and M. Terziovski, "The impact of operations and maintenance practices on power plant," *Journal of Manufacturing Technology Management*, vol. 25(8), pp. 1148–1173, 2014.

[2] V. Agarwal, K. A. Manjunatha, J. A. Smith, A. V. Gribok, and et al., *Machine Learning and Economic Models to Enable Risk-Informed Condition Based Maintenance of a Nuclear Plant Asset*. Idaho Falls, USA: INL/EXT-21-61984, Rev 0, Idaho National Laboratory, 2021.

[3] S. J. Remer, *Integrated Operations for Nuclear Business Operation Model Analysis and Industry Validation*. Idaho Falls, USA: INL/RPT-22-68671-Rev000, Rev 0, Idaho National Laboratory, 2022.

[4] *Projected Costs of Generating Electricity*. Paris: International Energy Agency, 2020.

[5] *Application of Wireless Technologies in Nuclear Power Plant Instrumentation and Control Systems*. Vienna: NR-T-3.29 Nuclear Energy Series, 2020.

[6] V. Agarwal, J. W. Buttles, L. H. Beaty, J. Naser, and B. P. Hallbert, "Wireless online position monitoring of manual valve types for plant configuration management in nuclear power plants," *IEEE Sensors Journal*, vol. 17(2), pp. 311–322, 2017.

[7] F. Calivá, F. S. De Ribeiro, A. Mylonakis, C. Demazi'ere, P. Vinai, G. Leontidis, and S. Kollias, "A deep learning approach to anomaly detection in nuclear reactors," in *2018 International Joint Conference on Neural Networks*, (Rio de Janeiro, Brazil), pp. 1–8, July 2018.

[8] N. Lybeck, K. D. Thomas, and C. Primer, *Plant Modernization Technical Program Plan for FY-2022*. Idaho Falls, USA: INL/EXT-22-28055, Rev 11, Idaho National Laboratory, 2022.

[9] Z. Ma, H. Bao, S. Zhang, M. Xian, and A. L. Mack, "Exploring advanced computational tools and techniques with artificial intelligence and machine learning in operating nuclear plants," *NUREG/CR-7294*, 2022.

[10] V. Agarwal, C. M. Walker, K. A. Manjunatha, T. J. Mortenson, N. J. Lybeck, A. C. Hall, R. A. Hill, and A. V. Gribok, *Technical Basis for Advanced Artificial Intelligence and Machine Learning Adoption in Nuclear Power Plants*. Idaho Falls, USA: INL/RPT-22-68942, Rev 0, Idaho National Laboratory, 2022.

[11] K. A. Manjunatha, V. Agarwal, and H. Palas, "Federated-transfer learning for scalable condition-based monitoring of nuclear power plant components," in *2022 Probabilistic Safety Assessment and Management*, (Honolulu, Hawaii), pp. 1–11, July 2022.

[12] K. A. Manjunatha and V. Agarwal, "Multi-kernel-based adaptive support vector machine for scalable predictive maintenance," in *2022 Annual Conference of the Prognostics and Health Management Society*, (Nashville, Tennessee), pp. 1–11, November 2022.

[13] K. D. Thomas, J. Remer, C. Primer, D. Bosnic, H. Butterworth, C. Rindahl, G. Foote, A. Droivoldsmo, G. Rindahl, R. McDonald, S. Lawrie, and E. Baker, *Analysis and Planning Framework for Nuclear Plant Transformation*. Idaho Falls, USA: INL/EXT-20-59537-Rev000, Rev 0, Idaho National Laboratory, 2020.

[14] White House Briefing Room, "President biden to catalyze global climate action through the major economies forum on energy and climate [fact sheet]." https://www.whitehouse.gov/briefing-room/statements-releases/2023/04/20/fact-sheet-president-biden-to-catalyze-global-climate-action-through-the-major-economies-forum-on-energy-and-climate/, 2023. Accessed: 27 JUL, 2023.

[15] A. Takyar, "AI use cases & applications across major industries." https://www.leewayhertz.com/ai-use-cases-and-applications, 2023. Accessed: 27 JUL, 2023.

[16] J. G. Kemeny, *The need for change, the legacy of TMI: Report of the President's Commission on the Accident at Three Mile Island*. Pergamon Press, 1979.

[17] Golden, "The profound promise of AI for the power sector." https://www.greenbiz.com/article/profound-promise-ai-power-sector, 2023. Accessed: 2023-07-07.

[18] A. Hall, J. C. Joe, T. M. Miyake, and R. L. Boring, "The evolution of the human systems and simulation laboratory in nuclear power research," *Nuclear Engineering and Technology*, vol. 55, no. 3, pp. 801–813, 2023.

[19] R. Awati, *Garbage in, garbage out*. TechTarget. https://www.techtarget.com/searchsoftwarequality/definition/garbage-in-garbage-out, (2023-06-01).

[20] M. Dainoff, L. Hettinger, A. C. Hall, J. H. Oxstrand, C. A. Primer, and J. C. Joe, *Using Systems Theoretic Process Analysis and Causal Analysis to Map and Manage Organizational Information to Enable Digitalization and Information Automation*. Idaho Falls, USA: INL/RPT-22-69058, Rev 0, Idaho National Laboratory, 2022.

[21] C. Alivisatos, *Evaluating Remote Operations for Advanced Nuclear Reactor Control: Feasibility, Benefits, and Implementation Criteria*. University of California, Berkeley, 2023.

[22] C. R. Kovesdi, Z. A. Spielman, J. D. Mohon, T. M. Miyake, R. A. Hill, and C. Pederson, *Development of an Assessment Methodology That Enables the Nuclear Industry to Evaluate Adoption of Advanced Automation*. Idaho Falls, USA: INL/EXT-21-64320-Rev00, Rev 0, Idaho National Laboratory, 2021.

[23] A. Y. Al Rashdan and S. W. St. Germain, *Automation of Data Collection Methods for Online Monitoring of Nuclear Power Plants*. Idaho Falls, USA: INL/EXT-18-51456, Rev. 0, Idaho National Laboratory, 2018.

[24] V. Agarwal, K. A. Manjunatha, A. V. Gribok, T. J. Mortenson, H. Bao, R. D. Reese, T. A. Ulrich, R. L. Boring, PhD, and H. Palas, *Scalable Technologies Achieving Risk-Informed Condition-Based Predictive Maintenance Enhancing the Economic Performance of Operating Nuclear Power Plants*. Idaho Falls, USA: INL/LTD-20-58848, Rev 0, Idaho National Laboratory, 2020.

[25] I. Huang, "Voicing for data engineering, the unsung hero.." https://towardsdatascience.com/voicing-for-data-engineering-the-unsung-hero-b91b6ef39dcd, 2020. Accessed: 27 JUL, 2023.

[26] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, p. 101805, 2023.

[27] Q. V. Liao, M. Pribić, J. Han, S. Miller, and D. Sow, "Question-driven design process for explainable AI user experiences," 2021.

[28] Consumer Financial Protection Bureau, "CFPB acts to protect the public from black-box credit models using complex algorithms [fact sheet]," 2022.

[29] D. A. Broniatowski, *Psychological foundations of explainability and interpretability in artificial intelligence*. Gaithersburg, MD: NIST Interagency/Internal Report (NISTIR)-8367, National Institute of Standards and Technology, 2021.

[30] Di Battista, A., Grayling, S., and Hasselaar, E., "Future of jobs report 2023," 2023. World Economic Forum, Geneva, Switzerland.

[31] S. LARSSON, "On the governance of artificial intelligence through ethics guidelines," *Asian Journal of Law and Society*, vol. 7, no. 3, p. 437–451, 2020.

[32] P. Cihon, M. M. Maas, and L. Kemp, "Should artificial intelligence governance be centralised? design lessons from history," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, (New York, NY, USA), p. 228–234, Association for Computing Machinery, 2020.

[33] A. Razzaque, *Artificial Intelligence and IT Governance: A Literature Review*, pp. 85–97. Cham: Springer International Publishing, 2021.

[34] S. Pickering and P. Davies, "Cyber security of nuclear power plants: Us and global perspectives," *Georgetown Journal of International Affairs, Jan*, vol. 22, 2021.

[35] K. Zetter, *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. Crown Publishers, 2014.

[36] A. O. Braseth, C. Nihlwing, H. Svengren, Ø. Veland, L. Hurlen, and J. Kvalem, "Lessons learned from halden project research on human system interfaces," *Nuclear Engineering and Technology*, vol. 41, no. 3, pp. 215–224, 2009.

[37] J. O'Hara, J. Higgins, S. Fleger, and P. Pieringer, "Human factors engineering program review model," *NUREG-0700 Rev. 3*, 7 2020.

[38] K. B. Bennett and J. M. Flach, *Display and Interface Design: Subtle Science, Exact Art*. USA: CRC Press, Inc., 1st ed., 2011.

[39] J. Rasmussen, *Mental models and the control of actions in complex environments*. Risø National Laboratory, 1987.

[40] K. J. Vicente and J. Rasmussen, "Ecological interface design: Theoretical foundations," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 4, pp. 589–606, 1992.

[41] J. M. O'Hara and S. Fleger, "Human-system interface design review guidelines," *NUREG-0711 Rev. 3*, 7 2020.

[42] A. Hossain and T. Zaman, "HMI design: An analysis of a good display for seamless integration between user understanding and automatic controls," in *2012 ASEE Annual Conference & Exposition*, no. 10.18260/1-2–21454, (San Antonio, Texas), ASEE Conferences, June 2012. https://peer.asee.org/21454.

[43] B. Hollifield, D. Oliver, I. Nimmo, and E. Habibi, *The high performance HMI handbook: A comprehensive guide to designing, implementing and maintaining effective HMIs for industrial plant operations.* Plant Automation Services, 2008.

[44] A. Duval, "Explainable artificial intelligence (XAI)," MA4K9 Scholarly Report, Mathematics Institute, The University of Warwick, 2019.

[45] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[46] M. Cubric, "Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study," *Technology in Society*, vol. 62, no. 4, 2020.

[47] C. Kovesdi, "Examining the use of the technology acceptance model for adoption of advanced digital technologies in nuclear power plants," in *Advances in Artificial Intelligence, Software and Systems Engineering* (T. Z. Ahram, W. Karwowski, and J. Kalra, eds.), (Cham), pp. 502–509, Springer International Publishing, 2021.

[48] J. Rasmussen, "Risk management in a dynamic society: a modelling problem," *Safety science*, vol. 27, no. 2-3, pp. 183–213, 1997.

[49] F. Davis, R. Bagozzi, and P. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Management Science*, vol. 35, no. 8, pp. 982–1003, 1989.

[50] N. Marangunić and A. Granić, "Technology acceptance model: A literature review from 1986 to 2013," *Universal Access in the Information Society*, vol. 14, pp. 81–95, March 2015.

[51] A. Vishwanath, "Impact of personality on technology adoption: An empirical model," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 8, pp. 803–811, 2005.

[52] S. Patel, "Digitalization is now a power sector imperative: Takeaways from connected plant conference 2023." https://www.powermag.com/digitalization-is-now-a-power-sector-imperative-takeaways-from-connected-plant-conference-2023/, 2023.

[53] "PSEG configuration baseline document for circulating water system," Tech. Rep. DE-CB.CW-0028(Z), Rev. 0, 1-1.

[54] NRC, "General Electric Systems Technology Manual: Chapter 11.1 - Circulating Water System, Rev. 09-11." Available at: `https://www.nrc.gov/docs/ML1125/ML11258A375.pdf`.

[55] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence program," *AI Magazine*, vol. 40, no. 2, p. 44, 2019.

[56] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[57] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[58] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[59] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[60] X. Wang and X. Tang, "Random sampling for subspace face recognition," *International Journal of Computer Vision*, vol. 70, pp. 91–104, 2006.

[61] L. Lin, P. Athe, P. Rouxelin, M. Avramova, A. Gupta, R. Youngblood, J. Lane, and N. Dinh, "Digital-twin-based improvements to diagnosis, prognosis, strategy assessment, and discrepancy checking in a nearly autonomous management and control system," *Annals of Nuclear Energy*, vol. 166, p. 108715, 2022.

[62] S. M. Abd Elrahman and A. Abraham, "A review of class imbalance problem," *Journal of Network and Innovative Computing*, vol. 1, no. 2013, pp. 332–340, 2013.

[63] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[64] L. Lin, A. Gurgen, and N. Dinh, "Development and assessment of prognosis digital twin in a NAMAC system," *Annals of Nuclear Energy*, vol. 179, p. 109439, 2022.

[65] D. Leslie, "Understanding artificial intelligence ethics and safety," *arXiv preprint arXiv:1906.05684*, 2019.

[66] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–158, 2012.

[67] W. J. Sutherland, A. S. Pullin, P. M. Dolman, and T. M. Knight, "The need for evidence-based conservation," *Trends in Ecology & Evolution*, vol. 19, no. 6, pp. 305–308, 2004.

[68] I. Chadès, E. McDonald-Madden, M. A. McCarthy, B. Wintle, M. Linkie, and H. P. Possingham, "When to stop managing or surveying cryptic threatened species," *Proceedings of the National Academy of Sciences*, vol. 105, no. 37, pp. 13936–13940, 2008.

[69] A. Zabeo, J. M. Keisler, D. Hristozov, A. Marcomini, and I. Linkov, "Value of information analysis for assessing risks and benefits of nanotechnology innovation," *Environmental Sciences Europe*, vol. 31, no. 1, pp. 1–8, 2019.

[70] H. Raiffa, "Decision analysis: introductory lectures on choices under uncertainty.," 1968.

[71] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "Review: A review of novelty detection," *Signal Processing*, vol. 99, p. 215–249, jun 2014.

[72] T. Wang, M. Cai, X. Ouyang, Z. Cao, T. Cai, X. Tan, and X. Lu, "Anomaly detection based on convex analysis: A survey," *Frontiers in Physics*, vol. 10, 2022.

[73] B. Subramanian, A. Paul, J. Kim, and K. W. A. Chee, "Metrics space and norm: Taxonomy to distance metrics," *Scientific Programming*, vol. 2022, OCT 6 2022.

[74] R. De Maesschalck, D. Jouan-Rimbaud, and D. Massart, "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[75] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008.

[76] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Database Theory — ICDT 2001* (J. Van den Bussche and V. Vianu, eds.), (Berlin, Heidelberg), pp. 420–434, Springer Berlin Heidelberg, 2001.

[77] NOAA National Centers for Environmental Information, "Monthly global climate report for annual 2022." `https://www.ncei.noaa.gov/access/monitoring/monthly-report/global/202213`, 2023. Accessed: 2023-02-19.

[78] O. Chornovol, G. Kondratenko, I. Sidenko, and Y. Kondratenko, "Intelligent forecasting system for npp's energy production," in *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pp. 102–107, 2020.

[79] T. Duan, A. Avati, D. Y. Ding, K. K. Thai, S. Basu, A. Ng, and A. Schuler, "Ngboost: Natural gradient boosting for probabilistic prediction," in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, JMLR.org, 2020.

[80] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, pp. 97–101, 2016.

[81] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[82] L. N. Tidjon and F. Khomh, "Never trust, always verify: a roadmap for trustworthy AI?," *arXiv preprint arXiv:2206.11981*, 2022.

[83] J. M. Wing, "Trustworthy AI," *Commun. ACM*, vol. 64, p. 64–71, sep 2021.

[84] B. Stanton and T. Jensen, "Trust and artificial intelligence," 2021-03-02 05:03:00 2021.

[85] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: a review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.

[86] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electron Markets*, vol. 31, pp. 447–464, 2021.

[87] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

APPENDIX A: BRIEFING REPORT

Briefing Paper

Author: Vivek Agarwal
Cody Walker
August 23, 2023

LWRS LIGHT WATER REACTOR SUSTAINABILITY

Explainable Artificial Intelligence Technology for Predictive Maintenance

Context of the Study

U.S. nuclear power plants have high operation and maintenance costs which make them noncompetitive in many energy markets. By leveraging artificial intelligence (AI) and machine learning (ML) these costs can be reduced, thus modernizing the process and improving their economics and reliability. However, to use AI/ML in nuclear power plants, it must be clear how and why the ML models reach their outcomes. This work focused on improving trust between operator and machine, one of the key aspects to improving the likelihood of nuclear power plants adopting AI/ML.

Current Barriers to Adopting Machine Learning

Several challenges to AI/ML technologies can be categorized (Figure 1) as historical, technical, business, regulatory and nuclear plant stakeholder readiness, and end-user acceptance / experience. Historical barriers encompass things such as instrumentation & control digitalization and modernization as well as nuclear safety events which have worked to limit the speed of technological advancement. Technological barriers include concerns of data quality, governance around the use of AI/ML, and cybersecurity. Regulatory readiness is the regulator trying to keep pace with the speed of technological innovations to allow for safe and secure use of AI. Stakeholders (nuclear plant owners, operators, and customers) readiness concerns the skills and knowledge required for plant personnel to successfully use AI.

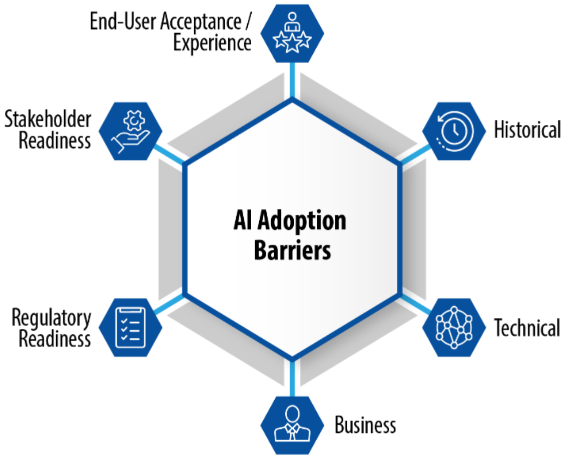


Figure 1. Reciprocal connectedness of barriers to AI/ML adoption in nuclear power.

Research-developed solutions

This research presents solutions on three aspects of AI technologies, i.e., performance, explainability, and trustworthiness (Figure 2) with specific metrics, user-centric visualization, and human-in-the-loop evaluation to build user confidence.

Explainable artificial intelligence approaches to improve human-machine trust.

To trust and use an ML model, operators need to know how and why it works. However, there is a trade-off between how well a model may perform and how readily explainable it is to the user. This trade-off has been explained in both qualitative and quantitative terms. Also, as new data arrive, they can be used to update the ML outcome and decision-making.

Trustworthiness

A copyrighted application has been created to put AI/ML tools into the hands of maintenance and diagnostic personnel. These users are assumed to have little to no ML experience, so the app has been tailored in such a way to explain how and why the ML model reached its conclusions. It also explains the historical context of the data and allows the user to arrive at their own conclusions, so they can keep a trust, but verify approach to plant monitoring.

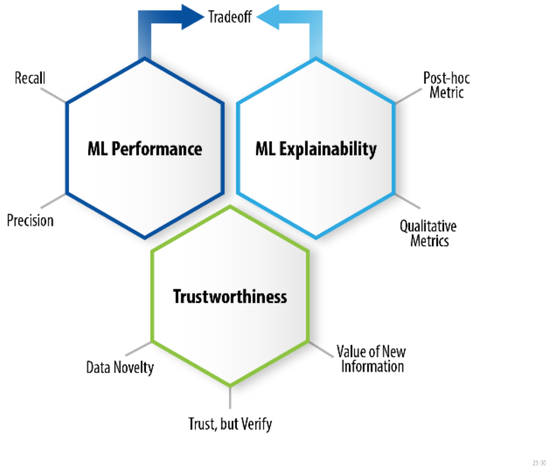


Figure 2. Aspects of AI technologies essential for decision-making.

User-centric visualization

AI/ML cannot be adopted without the human (called the "human-in-the-loop") who makes the ultimate decision on whether to schedule or perform a task, as shown in the research here, or perform another task. The developed application was tested with a large group of conference attendees to ascertain how the application appealed to a general audience of people with varying levels of ML experience. Overall, the app was received well, and their feedback will lead to improved versions of the app in the future.

Acknowledgments

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 Jason Tokey of DOE, and Bruce P. Hallbert and Craig A. Primer at Idaho National Laboratory (INL) for championing this effort. We thank Katie Stokes at INL for the technical editing of this report. We thank 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 Public Service Enterprise Group, Nuclear, LLC Monitoring and Diagnostic Center for their valuable technical contributions.

Contact

Vivek Agarwal | 765-631-1195 | vivek.agarwal@inl.gov

Cody Walker | 931-797-5403 | cody.walker@inl.gov

Craig A. Primer | 817-219-4363 | craig.primer@inl.gov

More on the LWRS Program: https://lwrs.inl.gov/

References

Explainable Artificial Intelligence Technology for Predictive Maintenance. INL/RPT-23-03620