Machine learning explanations by design: A Case study explaining the predicted degradation of a roto-dynamic pump.

Omnia Amin, Blair Brown, Bruce Stephen, and Stephen McArthur
Department of Electronic and Electrical Engineering, University of Strathclyde
Glasgow, G1 1RD, United Kingdom 0141 444 7027
omnia.amin@strath.ac.uk
blair.Brown@strath.ac.uk
bruce.Stephen@strath.ac.uk
s.mcarthur@strath.ac.uk

Valerie Livina
National Physical Laboratory (NPL)
Teddington, Hampton RD, TW110LW valerie.livina@npl.co.uk

Abstract

The field of explainable Artificial Intelligence (AI) has gained growing attention over the last few years due to the potential for making accurate data-based predictions of asset health. One of the current research aims in AI is to address challenges associated with adopting machine learning (ML) (i.e., data-driven) AI – that is, understanding how and why ML predictions are made. Despite ML models successfully providing accurate predictions in many applications, such as condition monitoring, there are still concerns about the transparency of the prediction-making process. Therefore, ensuring that the models used are explainable to human users is essential to build trust in the approaches proposed. Consequently, AI and ML practitioners need to be able to evaluate any available eXplainable AI (XAI) tools' suitability for their intended domain and end user while simultaneously being aware of the tools' limitations. This paper provides insight into various existing XAI approaches and their limitations to be considered by practitioners in condition monitoring applications during the design process for an ML-based prediction. The aim is to assist practitioners in engineering applications in building interpretable and explainable models intended for end users who wish to improve a system's reliability and help users make better-informed decisions based upon a predictive ML algorithm output. It also emphasizes the importance of explainability in AI. The paper applies some of these tools to an explainability use case in which real condition monitoring data is used to predict the degradation of a roto-dynamic pump. Additionally, potential avenues are explored to enhance the credibility of explanations generated by XAI tools in condition monitoring applications, aiming to offer more reliable explanations to domain experts.

1. Introduction

Recently, there has been extensive interest in eXplainable Artificial Intelligence [(1)] as researchers and practitioners in condition monitoring applications aim to make model

This is a peer-reviewed, accepted author manuscript of the following conference paper: Amin, O., Brown, B., Stephen, B., McArthur, S., & Livina, V. (2023). Machine learning explanations by design: a case study explaining the predicted degradation of a roto-dynamic pump. Paper presented at NDT 2023 - 60th Annual British Conference on NDT, Northampton, United Kingdom.

outputs more transparent and understandable [2]. Interpretability in the context of machine learning refers to the ability to explain or present information in a way that is understandable to practitioners [3]. However, for this paper, interpretability is used more broadly - currently, leading researchers [4] use interpretability to refer to directly transparent modelling mechanisms. Interpretability in this context is the level of understanding a human can have about the reason behind a decision made by a model. Additionally, interpretability is the degree to which a human can understand a model's behaviour and can accurately anticipate the model's results [4].

XAI refers to the methods and techniques used to understand and interpret the model or its predictions after training a model. A good explanation is when one can no longer ask why [5]. These definitions link interpretability to the machine learning process and provide an abstract objective for any machine learning explanation task.

Despite the recent increase in interest in XAI, much of the current research and practice relies on the researcher's subjective judgment of what constitutes a "good" explanation rather than a rigorous and objective standard [2].

In [6], three distinct groups of individuals who desire explanations with varying degrees of familiarity with ML have been identified, and each group has its reasons and requirements shown in Figure 1: Practitioners who have enough ML expertise to build and design ML and AI models; business stakeholders and regulators; and end users who are domain experts with no ML expertise. In Section 2, an important aspect to consider is how to match explainability tools to meet the specific needs of each individual or stakeholder involved in the engineering industrial application. This customization is crucial to ensure the explanations provided by XAI tools are tailored to the knowledge, context, and requirements of the recipients.

When beginning a machine learning project, it is best practice to consider the potential impact of the model on human beings or high-stakes decision-making. Maximum transparency is crucial in these cases to prevent fairness and security issues [7]. The lack of transparency and understanding in models severely impacts the trust and acceptance of these models, especially in high-stake sectors [7]. Additionally, using interpretable methods like explainable neural networks (XNN), gradient boosting machines (GBM), and Scalable Bayesian Rule (SBR) lists may not significantly decrease model accuracy. Starting with an interpretable model can also make debugging, explaining, and fairness auditing easier.

This paper aims to support practitioners in condition monitoring applications in creating interpretable models for end-users, aiming to improve the reliability of condition monitoring systems and empower users to make better decisions, underscoring the crucial role of explainability in AI. Some key considerations have been identified in this paper to help practitioners create explainable models. Using real data from condition monitoring of rotodynamic pumps, we have applied key factors identified to produce human-understandable explanations to domain experts in the field.

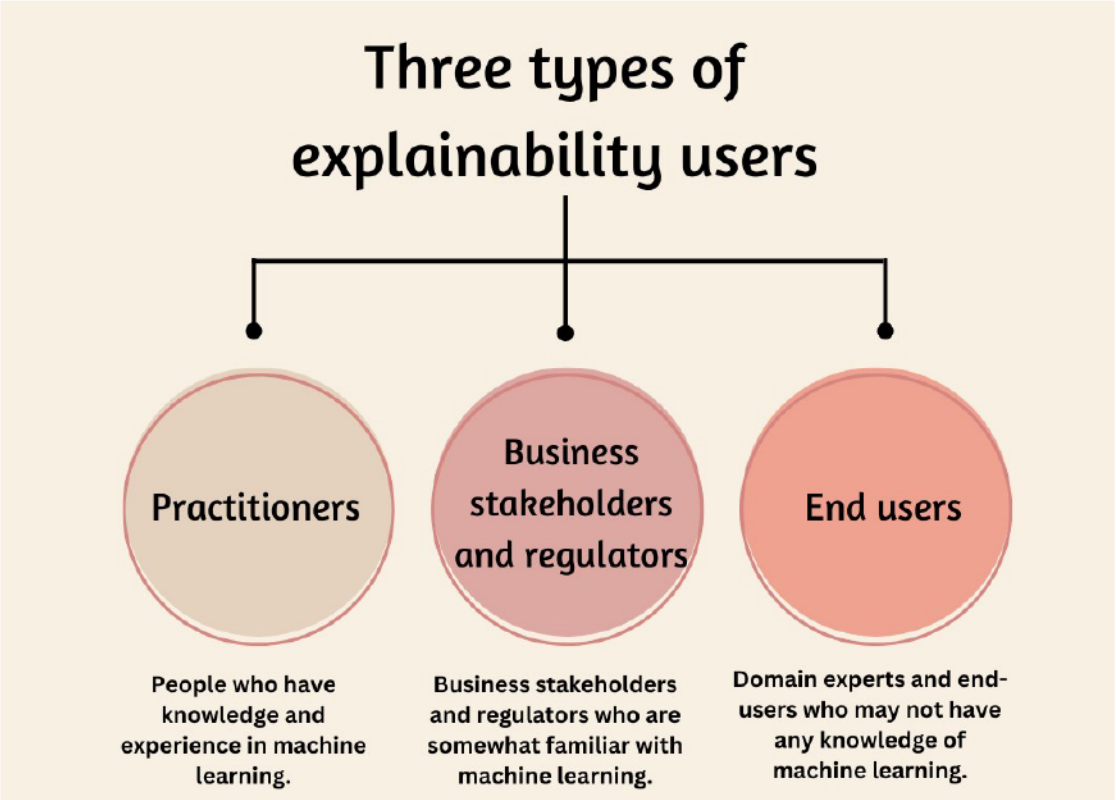


Figure 1. Types of explainability users.

2. Classification of explainability techniques

To ensure that the models built by practitioners are explainable to end-users, it is crucial for them to consider all possible approaches and types of explanations needed during the design process.
Different criteria for classifying XAI tools are discussed in this Section and shown in Figure 2.

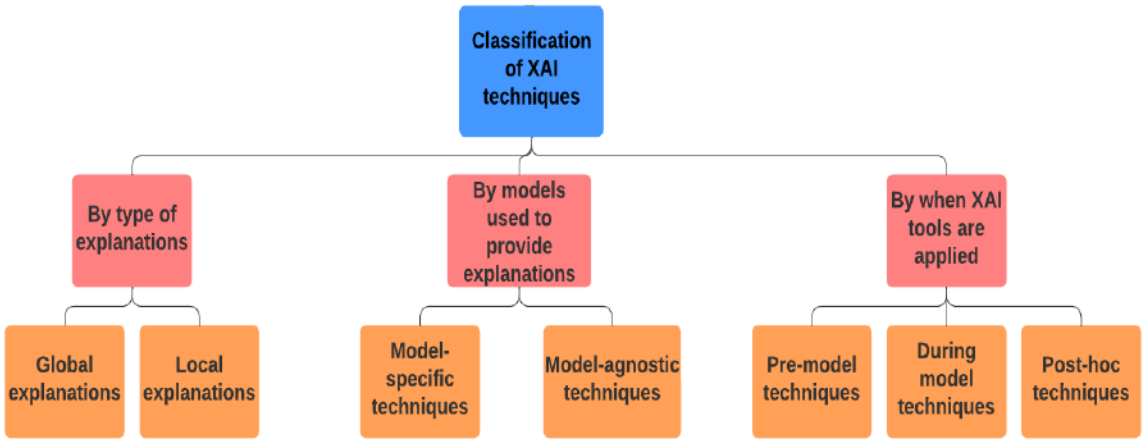


Figure 2. Classification of explainability techniques.

2.1 By type of explanations.

Evaluating a trained model from a global and local perspective is essential to understand its performance comprehensively. Global measures provide insight into the overall relationship between inputs and outputs but can sometimes be imprecise. On the other hand, local analysis examines specific rows or groups of similar data. Combining global and local interpretation techniques is often the most effective way to analyse a machine-learning model.

2.1.1 Global explanations

Global explanations assess the magnitude of the impact or identify the features that have the most significant impact on the model's accuracy. Such global explanations are helpful in making informed decisions or supporting or dismissing the idea that a specific feature is substantial [(9)].

2.1.2 Local explanations

Local explanations outline the reasoning behind a specific prediction made by an ML model. They focus on providing insight into segments of the relationship between the inputs and outputs of a machine learning model, such as certain groups of input records and their corresponding predictions, specific ranges of predictions and the input data associated with them, and even individual data points. These explanations are valuable for gaining an in-depth understanding or identifying problems and can be used to understand specific predictions. A computer vision task may consist of selecting pixels that impacted the classification most [(10)].

2.2 By method used to provide explanations.

Another vital way to classify explanation techniques depends on the model being used. These techniques can be model-agnostic or model-specific. Model-agnostic methods can be applied to any model, while model-specific methods are limited to specific models.

2.2.1 Model-Specific

Model-specific methods involve examining or utilizing the internal components of a model. Model-specific interpretability methods generally use the model itself for interpretation, which can provide more accurate information.

2.2.2 Model-agnostic

Model-agnostic methods examine the connection between the input and output of trained models without relying on their internal structure [29]. These methods are beneficial when no theory or understanding of what is happening within the model exists. Model-agnostic explanation methods typically only have access to the model's output and do not require any information about the model's internals, making them widely applicable and flexible. These methods have the benefit of not affecting the model's performance as they are separate from the internal functioning of the model by treating the machine learning models as black-box functions [11].

Popular examples of model-agnostic post-hoc explanations methods include Local Interpretable Model-agnostic Explanations (LIME), Layer-wise Relevance Propagation (LRP), and Shapley Additive Explanations (SHAP).

While model-agnostic interpretability techniques are convenient and have advantages, they often rely on substitute models or other estimates that can reduce the precision of the information they provide. In contrast, model-specific interpretability techniques use the model to be explained directly, resulting in more accurate measurements [12].

2.3 By when XAI tools are applied.

Explainability techniques can be classified based on when they are used within the modelling process. These techniques are commonly grouped into three categories: pre-model explanations, during-model explanations, and post-hoc explanations. Post-hoc techniques are considered the most common because they offer flexibility and can be applied to existing trained models.

2.3.1 Pre-model techniques

Pre-model techniques refer to the interpretability methods used before selecting and developing models; they are closely related to exploratory data analysis [13].

Exploratory data analysis is summarizing the main characteristics of a dataset to gain a deeper understanding of the dataset, including calculating statistics such as the mean, standard deviation, range, missing samples, and more. However, it's important to note that statistical features alone are insufficient to analyse the data thoroughly [8].

2.3.2 During model techniques

Interpretability can be easily achieved by using a limited set of algorithms that produce interpretable models. Linear regression, logistic regression, and decision trees are examples of commonly used interpretable models [4].

- *Linear regression model.*
  Linear regression is a model that predicts the target by taking the weighted sum of the feature inputs. The linear relationship between the inputs and the target makes it easy to interpret the results. Linear regression models are widely used by statisticians, computer scientists, and other professionals who deal with quantitative problems.

- *Logistic regression.*
  When dealing with classification problems where the objective is to determine the probability of data belonging to a particular class, there are more suitable approaches than linear regression; in such cases, logistic regression is a better alternative as it models the probability of a discrete outcome based on input variables. Unlike linear regression, which focuses on linear interpolation between data points [8]. Logistic regression involves creating a model to predict the probability of a discrete outcome based on an input variable [13].

- *Decision trees.*
  Decision trees are another example of interpretable models. Decision trees facilitate prediction and classification while possessing inherent explainability [8]. They function by recursively dividing data based on their attributes and grouping them into subsets in a hierarchical structure [15].

2.3.3 Post-model techniques

Post-hoc explanation techniques are used to explain how models that are not interpretable by design work. These techniques are applied to a model after it has been developed to understand how predictions are made. Typically, post-hoc techniques involve using an external, simpler model that mimics the behaviour of the complex model to make it more transparent. Post-hoc interpretability is a different way of gaining insight from trained models. Although post-hoc interpretation may not provide a detailed understanding of how a model works, it provides valuable information for practitioners and users of machine learning [16] [17].

Visual explanation tools provide visual insight into how ML models make their predictions. Visualization techniques are produced to help make the decision-making process more understandable to general users who can easily interpret the visual explanations. Some examples of visually explainable AI tools include LRP, saliency maps, Class Activation Maps (CAM), and Gradient-weighted Class Activation Maps (Grad-CAM). LRP [18] is an attribution method that evaluates the degree to which an input (or weight) affects the prediction to determine the most important features. Saliency maps [28] is a heatmap visualization tool that gives important factors to each pixel in an image and shows its contribution to the prediction. CAM and Grad CAM are similar explainability tools best used with convolutional networks. They produce heatmaps

visualizations that highlight pixels or regions of an image the model uses to make its predictions. Grad CAM [19] is an advanced version of CAM that uses the gradients of the model's output to the input image to create the heatmap visualization.

Individual Conditional Expectation (ICE) and Partial Dependence (PD) [4] plots can provide insights into the relationship between a specific feature and the outcome of a model, such as whether it is monotonic or linear. However, it should be noted that the average effect of a feature on the model's decision can be misleading and can obscure important interactions among variables. Therefore, it is a better approach when both ICE and PD plots are used, as they complement each other in identifying the relationships between variables. Accumulated Local Effects (ALE) are similar to PD in explaining how features impact the predictions of a model on average. However, ALE has an advantage over PD as it addresses the bias that can arise when the feature of interest is highly correlated with other features, which is not considered in PD [4].

Shapley Additive Explanations (SHAP) [9] is a post-hoc method used to explain the prediction of an instance by determining the contribution of each feature to the output. SHAP is most used for tabular data. SHAP uses Shapley values to assess the effect of individual features on the model's outcome. SHAP guarantees consistency and local accuracy by considering all possible predictions and input combinations. It comes in two variants: KernelSHAP and TreeSHAP. Kernel SHAP is a model-agnostic method based on LIME and Shapley values but has high computational complexity. Tree SHAP [26] computes exact SHAP values for decision tree-based models. Asymmetric Shapley Values (ASV) [20] is a variation of SHAP that considers the causal relationship between variables in the model explanation process. ASV values are asymmetric, unlike Shapley values, and they are often used in model fairness analysis because they can capture the indirect effects of a variable on the model.

LIME, [10] a model-agnostic method, has seen many successful applications in different domains. LIME uses the perturbation of a single data sample and with the help of a surrogate model considers the impact of the perturbed data sample in the predictions. LIME relies on a surrogate model to indirectly solve the explanation problem, and the quality of the explanation largely depends on the quality of the surrogate fit. This can result in high computational costs and uncertainty, leading to variable explanations for the same input sample.

"Anchors" [10] and the related concepts explain individual predictions of any black-box classification model by finding a decision rule that "anchors" the prediction. The resulting explanations are IF-THEN statements, which define regions in the feature space where the predictions are fixed to the class of the data point being explained [30]. This ensures the classification remains the same, regardless of changes to other feature values that are not part of the anchor. The main drawback of Anchors is that it only supports textual and tabular data. It generates explanations in tables and text, but it is limited to those types of data.

Example-based explanation or explanation-by-example techniques are other examples of post-hoc explainability approaches which involve selecting instances from testing and training datasets that the model predicts well and generates global or local explanations for the task model. Unlike other techniques, example-based techniques do not highlight the most important features, but instead, they create a simplified understanding of the black-box model [26]. There are several example-based explanation techniques, such as

contrastive explanations (CEM), k-nearest neighbour (kNN), trust scores (TSM), and counterfactuals. CEM [21] provides local explanations for predictions by comparing them to another prediction. CEM is based on the concepts of pertinent positives and negatives. Pertinent positives are the minimal features that can justify the prediction. In contrast, relevant negatives are the minimal set of features that must be absent from asserting the prediction, and both concepts provide a complete explanation.
KNN can be used as an example-based tool to provide local explanations for black-box classifiers by identifying the most common among the k-closest data samples.
TSM is classified as an explainability tool. It is a way to determine the confidence of a black box classifier model in making a specific prediction. However, it does not explain the prediction but generates a trust score. It uses a modified version of the nearest neighbour classifier to calculate a trust score for a specific instance.
Finally, counterfactuals [22] involve thinking about what changes could have been made to an input to produce a different prediction. It generates local explanations by recognizing the modifications to be made to an input instance to produce a different outcome.

2.4 The XAI classification range

XAI is a wide-ranging field that encompasses a variety of techniques and definitions. The field itself is new and rapidly evolving, and as a result, there are many different approaches and interpretations of explainability. The concept of transparent models or interpretable models, which relates to understanding the internal mechanisms of an AI system and how it reaches decisions, sits at one end of the XAI spectrum. On the opposite end sits post-hoc explainability techniques, which concentrate on the ability to explain the decisions made by black-box systems. In between these two extremes, various methods and approaches fall under the XAI umbrella, such as those previously mentioned in this paper. These different approaches have one common goal but different trade-offs. The definition of XAI can change depending on the context, stakeholders involved, and application. The overall goal of XAI is to enhance the trust and accountability of AI systems by enabling a better understanding of how they work and why they make certain decisions. The previous text discusses the topic of explainability and references various publications on the subject. The authors support using explainable AI to increase the trustworthiness and transparency of AI systems by providing additional information in the form of explanations that gives insight into its decision-making or internal workings.

3. Considerations when designing an explainable ML methodology.

Selecting the right XAI approach for your application can be challenging due to the various available options. However, certain key factors must be considered while designing an explainable ML system. These are considered below and depicted in Figure 3.

1. Defining the application goal(s) is crucial before starting the design process. Applying XAI can include enhancing the transparency of the ML model; being compliant with AI regulatory requirements; identifying potential bias in the decision-making process; and helping naïve users understand how ML models work.

2. Understanding data is one of the most critical factors in designing any ML methodology and choosing the right XAI tool. Data can help reveal potential biases that the ML model may have and can help limit choices of XAI tools. If we have structured data, we can look for XAI tools to handle this kind of data. Some options would be LIME and SHAP. However, if we have unstructured data like text and images, we can investigate other XAI tools such as saliency maps or gradient-based XAI tools. Additionally, developers can validate XAI explanations generated by identifying relationships and patterns in the data using pre-model XAI tools.
3. Machine learning model selection: as mentioned in Section 2, some explainability tools are model-specific while others are model-agnostic. A compatible XAI tool with the ML model used should be chosen. Choosing the appropriate XAI tool for the ML model is important to ensure that the explanations generated are accurate and meaningful.
4. Meeting the stakeholders' unique needs by considering the level of technical skills they have in the design process will help build an effective XAI methodology that is both explainable and trustworthy for all users involved. To attain explainability in AI systems, including human-centred design principles in the development process is crucial. This involves designing AI systems that consider the user's requirements and preferences through user research, user testing, and iterative design processes. As previously mentioned in Section 2. Different XAI tools require a level of expertise in AI. What is considered easy for practitioners might not be comprehensible for regulators or end users. Explanations generated from the XAI tool should be tailored to the user's needs [23], some examples now follow. Business stakeholders pose broader, business-oriented inquiries when seeking an explanation; that's why they frequently favour a layman's explanation dto obtain information that fosters confidence in the model's expected performance in the intended context. However, regulators aim to verify that the model complies with a particular set of standards. Domain experts have advanced knowledge of the environment in which the model operates and usually, they have no ML background. They seek explanations as a tool to help in the decision-making process [7].
5. Determine the level of explainability needed for the application. The level of explainability required depends on the use case, the stakeholders involved, and the regulatory requirements. Generally, trying out simple and transparent models is the first step toward designing an explainable ML model. If the performance of these simple models is acceptable to the end users, then there is no need to use complex models. However, if the performance is compromised, complex models can be used and the appropriate XAI technique can be used to explain these models.
6. Determine the evaluation measures to validate the XAI explanations generated. Evaluating the quality of XAI explanations for specific use cases is important to enhance stakeholders' trust in ML models. Different measures can be used depending on the end users' needs. These measures can be the robustness of explanations generated, using various XAI tools and comparing their explanations, getting help from domain experts to evaluate explanations generated, and providing feedback. Finally, testing or altering features used and observing the effect on the ML model's performance.

4. Modelling Trade-offs

Developing machine learning models requires careful consideration of various trade-offs, including accuracy, security, performance, and explainability [23]. Usually, highly accurate models are more complex and can be more challenging to explain. Also, models that are highly explainable and transparent may make it easier to identify potential security or privacy vulnerabilities. However, this increased transparency may also make the model more vulnerable to attacks by malicious actors seeking to exploit these weaknesses. The choice between these trade-offs can be done through understanding data, the specific needs of the application and the priorities of the end-users and stakeholders.

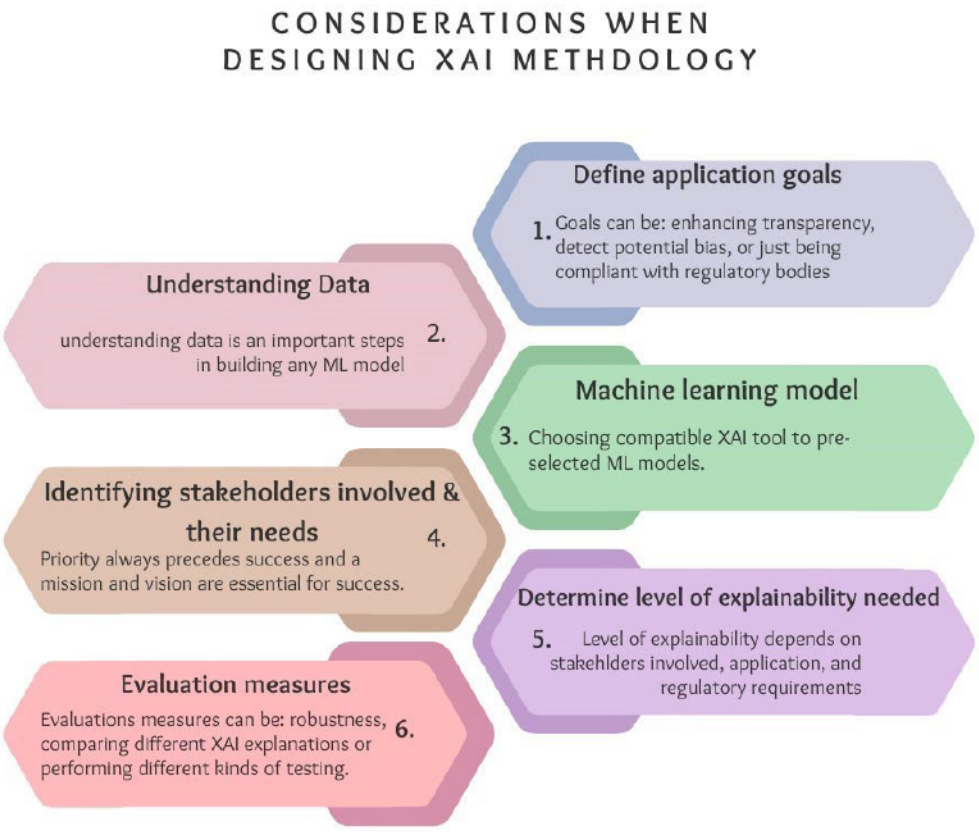


Figure 3. Considerations when designing XAI methodology.

5. Connecting the dots: matching end users' needs with explainable AI tools.

One of the significant goals in explainability is getting stakeholders involved, ensuring they understand the explanations generated, and therefore developing trust ML models used [26]. End users from different domains have varying preferences for how information is presented to them. Previous publications [24] [25] shed light on users' needs and level of expertise with regard to the choice of appropriate explanations. Choosing the best explanation for a specific model is essential as it may be required to be understood by a wide range of end-users, including non-technical users, data scientists, and domain

experts, each with unique perspectives and potential regulatory requirements. Users' understandability will increase if the explanations are presented in an appropriate form. Explainability tools mentioned in Section 2 usually provide different explanations, including analytical, visual, rule-based, and textual. Depending on the end user(s), an appropriate type of explanation should be provided. Tailoring the type of explanation provided to end-users is crucial, as different formats may be more suitable for different individuals. For instance, domain experts who are non-technical users may benefit from visuals and storytelling such as rule-based, whereas text-based explanations may be more appropriate for naïve users.
Providing too much information can lead to confusion and decrease the understanding and trust of users. The need and type of explanation required varies depending on the context, such as the user, their role, prior knowledge, and the application. As a specific example, nuclear plant operators require a different explanation to a nuclear regulator - explanations should hence consider all user's needs, roles, and goals.

6. A framework for non-ML expert explanations.

Many XAI tools currently in use generate explanations that are tailored for individuals with a background in ML. However, it can be difficult to effectively communicate these explanations to stakeholders with less technical expertise, as there is often a significant gap in knowledge between ML experts and non-experts. Using all the considerations in Section 2, the authors have previously proposed a new framework [31] to give non-technical users a deeper understanding of how ML models work and provide them with explanations that are both easy to understand and casually link the XAI-generate explanations. The framework aims to generate explanations that are both easily understood by humans and can be presented graphically for the benefit of non-experts in ML. The goal is to create multiple methods for extracting explainability from tools that make predictions or provide diagnostic information. To achieve this goal, the authors have proposed a framework comprising five interconnected stages for giving explanations.
The framework generates text-based and visual-based explanations to help bridge the knowledge gap between ML experts and non-experts.

7. Select a case study.

Real data from feed-water pump condition monitoring was used to predict the movement of thrust bearings. This movement is analogous to degradation as movement leads to wear on the associated pads, reducing the pump's lifespan. The ML model implemented in this study uses predetermined operational profiles, such as flow and head, to predict the median thrust-bearing position. The goal of this case study is to consider all the key factors mentioned in Section 2 before starting the design process.

7.1 Define application goals.

We aim to understand the impact of two different operating profiles on the degradation of a pump. The operating profile of a pump refers to the specific conditions under which it

is used, such as the flow rate, pressure, and thrust-bearing movement. By analysing different operating profiles, the study aims to identify which conditions lead to the most degradation on the pump and how this degradation will progress over time.
The second goal is to present the findings in a way that is easy to understand for the domain expert who represents the end user in this application. This includes presenting the explanations in a way that is easy to interpret. The goal is to make the study result accessible and actionable for the end user to make informed decisions about operating and maintaining the pump.

7.2 Understanding the Data

The first step in the design process is understanding the data we have. In this case study, we have structured tables representing the input data, so we have limited our choices of XAI tools to those that can work with structured data. The study uses two data sets: one from normal operation and one from an operational state where the associated power plant is refuelling, hence representing different pump operating conditions. The data is evaluated for quality and then transformed into a format that can be utilized by machine learning or analytic models. This application uses multivariate time series data for flow and head to predict thrust-bearing movements.

7.3 Select machine learning models.

Three simple models have been used, and their performances have been evaluated: linear regression, random forest, and XGboost. These models were used to predict the positions of thrust-bearing movements. After the evaluation, it was found that their performances were within acceptable limits, so there was no need to try out complex models. Despite their simplicity, these machine learning models can be challenging to understand for non-experts.

7.4 Identify stakeholders involved.

When identifying stakeholders involved in ML applications, several groups may be impacted by or have a stake in the outcome. All relevant stakeholders must be included to address potential ethical, legal, or business implications and ensure the model is fair, transparent, and accurate. Some of the key stakeholders involved in this case study are AI and ML developers who are responsible for building an explainable ML model that domain experts will use to predict thrust-bearing wear movements, domain experts who will use this ML model and need human-understandable explanations of the predictions made by this model, and regulators in the nuclear industry who will need to review and approve these models before deploying them to ensure they comply with regulations.

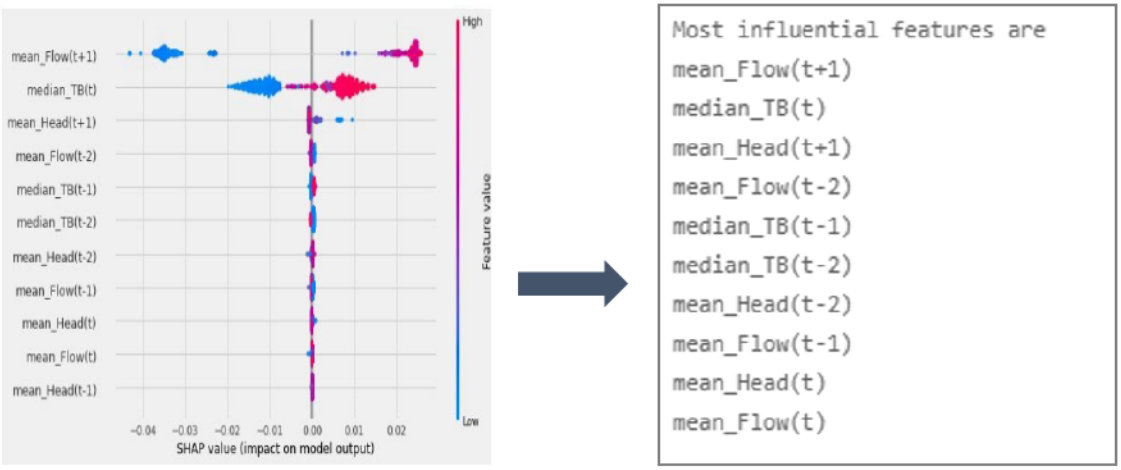
7.5 Determine the level of explainability needed.

As mentioned in Section 5, The level of explainability required depends on the use case, the stakeholders involved, and the regulatory requirements. For this case study, XAI

explanations will be presented to domain experts who are not ML experts. Therefore, explanations must be in a human-understandable context.
After choosing the appropriate models, we selected SHAP to explain the predictions generated from these ML models. SHAP offers a robust theoretical foundation, delivering both local and global explanations, can handle structured data in tables, and outperforms other explainability tools which can handle this kind of data in terms of reliability. However, SHAP provides explanations in the form of complex plots that are difficult to understand for users without ML expertise. To deliver human-understandable explanations, algorithms have been developed to translate SHAP technical explanations and generate text-based and graphical-based explanations for domain experts to help them understand the explanations generated.

7.6 Case-study output

Following all the considerations outlined in Section 2, we design an explainable ML model targeting domain experts with no ML background. An example of explanations generated in text format is shown in Figure 4, aimed at providing a simpler and quicker understanding for individuals without expertise in machine learning. Figure 4 displays an example of explanations generated by SHAP, showing the most important features and text-based explanations corresponding to the SHAP plot [31].
In Figure 5 [31] an example of a text-based explanation corresponding to the SHAP force plot is shown in the Figure. In Figure 5, we can observe that the most important feature for this specific instance is mean_Flow(t+1) which positively impacts the output, driving the output to increase. On the other hand, mean_Head(t+1) has a negative impact on the output deriving the output to decrease.



SHAP global explanation. Text-based explanation.

Figure 4. SHAP global explanation and the corresponding text-based explanation.


```
  instance no 0  taken at time:  30/10/2015 14:04
features affecting the median_TB(t+1) in order are
  mean_Flow(t+1)
  median_TB(t)
  mean_Head(t+1)
  mean_Flow(t-2)
  median_TB(t-2)
  median_TB(t-1)
  mean_Head(t-2)
  mean_Flow(t-1)
  mean_Head(t)
  mean_Flow(t)
  mean_Head(t-1)
impact of each feature value on output
mean_Flow(t-2)   479.3611552 High value pushes median_TB(t+1) lower
mean_Head(t-2)   193.4650223 High value pushes median_TB(t+1) higher
median_TB(t-2)   0.001806837 High value pushes median_TB(t+1) lower
mean_Flow(t-1)   478.8041306 High value pushes median_TB(t+1) lower
mean_Head(t-1)   193.0229993 High value pushes median_TB(t+1) higher
median_TB(t-1)   -3.11e-06 High value pushes median_TB(t+1) higher
mean_Flow(t)   477.8979441 High value pushes median_TB(t+1) higher
mean_Head(t)   192.6907872 Low value pushes median_TB(t+1) lower
median_TB(t)   -0.000458766 High value pushes median_TB(t+1) higher
mean_Flow(t+1)   479.9311702 High value pushes median_TB(t+1) higher
mean_Head(t+1)   193.4374611 High value pushes median_TB(t+1) lower
```

Figure 5. SHAP force plot and the corresponding text-based explanation.

The proposed approach transforms the explanations derived from the SHAP technique into a format that is easily understandable for nuclear domain experts who usually do not have enough ML expertise. This improved interpretability enhances trust and understanding of AI systems, facilitating their integration into nuclear industry where transparency and reliability are crucial.

Future directions of development

XAI tools aim to create new and improved ways of explaining and examining AI and ML models. Building on the research presented in this paper, the authors are currently working on creating evaluation measures to validate explanations generated by XAI tools. These explanations may not align with the knowledge and expertise of domain experts who deeply understand the problem at hand. In that case, their reliability and validity may be called into question. The research is working on validating XAI explanations against domain knowledge to ensure that the explanations provided are reliable and valid [28]. This validation process involves comparing the XAI-generated explanations with the domain knowledge to see if they align with what is already known about the problem. Consistency between the explanations and domain knowledge can increase confidence in the accuracy of the explanations and the underlying AI models.

On the other hand, inconsistencies may signal potential problems or limitations in the AI models or the XAI tools themselves. By validating XAI-generated explanations against

domain knowledge, researchers can establish the reliability and validity of these tools and strengthen trust in the outcomes produced by AI models. Furthermore, this validation process can help identify gaps or inaccuracies in domain knowledge, which can drive further research and investigation to improve our understanding of the problem.
Authors are also looking into the possibility of automatically creating graph-based representations of the text-based explanations generated from the XAI tool. A knowledge graph represents information that connects entities and their relationships in a structured format. Creating knowledge graphs will provide a more intuitive and interactive representation of the information in the generated XAI explanations. This representation will give the users machine-readable information about the domain. We will also validate this knowledge graph by incorporating domain knowledge into the XAI process.

Discussion and conclusions

Given the wide range of options and definitions of XAI, practitioners in condition monitoring applications have the challenging task of choosing the right XAI tool for their applications. In this paper, some key factors identified in Section 2 were considered during designing an explainable system have been discussed. There are many explainable AI tools; here, we have summarized a guide for ML and AI practitioners to follow in choosing the appropriate explainable AI tool for their application.
In the context of condition monitoring systems, the application of the approach discussed in this paper can provide valuable insights and explanations to aid practitioners in designing explainable systems. Utilizing real condition monitoring data from a rotodynamic pump, and following the approach identified, human-understandable explanations have been generated to aid stakeholders understand the explanations generated.
In summary, when choosing an XAI tool, developers should fully understand the application goal, understand the data they use, choose a compatible XAI tool for their ML model, identify all the stakeholder involved and their needs, determine the level of explainability needed, and finally have an evaluation measure that they can use to validate the explanations generated.
XAI offers significant advantages to various application domains that use AI-based systems. These applications should consider the needs of end users to help them engage with the explanations generated.

Acknowledgements

The authors gratefully acknowledge the support from the UK's National Physical Laboratory (NPL) and the Advanced Nuclear Research Centre (ANRC) at the University of Strathclyde, Glasgow.

References

1. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: https://doi.org/10.1109/access.2018.2870052.
2. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019, doi: https://doi.org/10.1016/j.artint.2018.07.007.
3. Finale Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv: Machine Learning*, 2017.
4. Christoph Molnar, "Interpretable Machine Learning," *Github.io*, Aug. 27, 2019. https://christophm.github.io/interpretable-ml-book/
   L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, and Lalana Kagal, "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning," *ArXiv*, 2018.
5. M Munn, and D Pitman, "Explainable AI for Practitioners: Designing and implementing explainable ML solutions." 2022.
6. S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward," Humanities and Social Sciences Communications, vol. 7, no. 1, pp. 1–7, Jun. 2020, doi: https://doi.org/10.1057/s41599-020-0501-9.
7. U. Kamath, and J. Liu," Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, Springer International Publishing". (2021).
8. S Lundberg and M., G Erion., H Chen. et al. "From local explanations to global understanding with explainable AI for trees". Nat Mach Intell 2, 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9
9. M Tulio Ribeiro, S Singh, and C. Guestrin. " Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016).
10. A Holzinger,et al. (2022). xxAI – " xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. A. Holzinger, R. Goebel, R. Fong et al. Cham, Springer International Publishing: 3-10.
11. P Hall and N Gill." An introduction to machine learning interpretability." O'Reilly Media, Incorporated. (2019).
12. J.W. Tukey, Exploratory Data Analysis Michael Waskom and the seaborn development team. mwaskom/seaborn Version (Addison-Wesley, Reading, 1977)
13. M Mehta, V Palade,& I Chatterjee, (Eds.). Explainable AI: Foundations, Methodologies and Applications (Vol. 232). Springer Nature. (2022).
14. YY Song, Y Lu Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015;27(2):130-135. do i: 10.11919/j.issn.1002-0829.215044
15. Z. C. Lipton, "The Mythos of Model Interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018, doi: https://doi.org/10.1145/3236386.3241340.

16. A Holzinger, C. Biemann, , Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? No. Ml, pp. 1–28 (2017). http://arxiv.org/abs/1712.09923
17. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. Müller, "Layer-Wise Relevance Propagation: An Overview," 2019.
18. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: https://doi.org/10.1007/s11263-019-01228-7.
19. C. Frye, C. Rowat, and I. Feige, "Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability," *arXiv.org*, Dec. 20, 2021.
20. A Jacovi, S Swayamdipta, S Ravfogel, Y Elazar, Y Choi, and Y Goldberg. 2021. Contrastive Explanations for Model Interpretability. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
21. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electronic Journal*, 2017, doi: https://doi.org/10.2139/ssrn.3063289.
22. C. Mougan, G. Kanellos, and T. Gottron, "Desiderata for Explainable AI in statistical production systems of the European Central Bank," *arXiv.org*, Feb. 12, 2022. https://arxiv.org/abs/2107.08045.
23. I. kobrinska. etl ,Principles and Practice of Knowledge Discovery in Databases, Cham, Springer International Publishing.
24. I. Nunes and D. Jannach, "A systematic review and taxonomy of explanations in decision support and recommender systems," *User Modeling and User-Adapted Interaction*, vol. 27, no. 3–5, pp. 393–444, Oct. 2017, doi: https://doi.org/10.1007/s11257-017-9195-0.
25. G. Ras, N. Xie, M. Van Gerven, and D. Doran, "Explainable Deep Learning: A Field Guide for the Uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–397, Jan. 2022, doi: https://doi.org/10.1613/jair.1.13200.
26. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: https://doi.org/10.3390/e23010018.
27. U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, and D. A. Keim, "Towards a Rigorous Evaluation of XAI Methods on Time Series," *arXiv.org*, Sep. 17, 2019. https://arxiv.org/abs/1909.07082 (accessed May 25, 2023).
28. M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," *arXiv:1606.05386 [cs, stat]*, Jun. 2016, Available: https://arxiv.org/abs/1606.05386
29. M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision ModelAgnostic Explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, Available: https://ojs.aaai.org/index.php/AAAI/article/view/11491
30. O. Amin, B. Brown, B. Stephen, and S. McArthur, "A Case-study Led Investigation of Explainable AI (XAI) to Support Deployment of Prognostics in

the industry," *PHM Society European Conference*, vol. 7, no. 1, pp. 9–20, Jun. 2022, doi: https://doi.org/10.36001/phme.2022.v7i1.3336.