A Case-study Led Investigation of Explainable AI (XAI) to Support Deployment of Prognostics in industry

Omnia Amin[1],Blair Brown[2],Bruce Stephen[3] and Stephen McArthur[4].

[1,2,3,4] *Department of Electronic and Electrical Engineering, University of Strathclyde ,250 George St, Glasgow G1 1XQ*
omnia.amin@strath.ac.uk
Blair.Brown@strath.ac.uk
Bruce.Stephen@strath.ac.uk
S.mcarthur@strath.ac.uk

ABSTRACT

Civil nuclear generation plant must maximise it's operational uptime in order to maintain it's viability. With aging plant and heavily regulated operating constraints, monitoring is commonplace, but identifying health indicators to pre-empt disruptive faults is challenging owing to the volumes of data involved. Machine learning (ML) models are increasingly deployed in prognostics and health management (PHM) systems in various industrial applications, however, many of these are black box models that provide good performance but little or no insight into how predictions are reached. In nuclear generation, there is significant regulatory oversight and therefore a necessity to explain decisions based on outputs from predictive models. These explanations can then enable stakeholders to trust these outputs, satisfy regulatory bodies and subsequently make more effective operational decisions. How ML model outputs convey explanations to stakeholders is important, so these explanations must be in human (and technical domain related) understandable terms. Consequently, stakeholders can rapidly interpret, then trust predictions better, and will be able to act on them more effectively. The main contributions of this paper are: 1. introduce XAI into the PHM of industrial assets and provide a novel set of algorithms that translate the explanations produced by SHAP to text-based human-interpretable explanations; and, 2. consider the context of these explanations as intended for application to prognostics of critical assets in industrial applications. The use of XAI will not only help in understanding how these ML models work, but also describe the most important features contributing to predicted degradation of the nuclear generation asset.

1. INTRODUCTION

Although there are many different approaches in PHM, AI and ML powered techniques have recently seen a surge across applications in different industries. These Industries are continuously exploring AI and ML methods to ensure reliable and sustainable operations for their industrial assets. The goal of using these techniques is to carefully maintain industrial assets, to ensure that they fulfil their dedicated functions and also to avoid any unnecessary asset downtime. However, in industries where safety and reliability are crucial, the use of AI techniques impose a challenge of non-transparency to stakeholders. Stakeholders need to understand how ML techniques work and how they produce their outputs in order to build trust in decisions based upon these outputs and realise AI/ML deployments within their industries. Explainable AI (XAI) helps in explaining these techniques and make it more transparent to stakeholders. XAI has a vital role in PHM systems as it helps nurture confidence in AI techniques used while the function and performance of the underpinning AI systems and the associated asset remain intact. This paper illustrates the need for XAI in PHM and how XAI can help non-ML experts adopt ML models through demonstration on diagnostic and anomaly detection case studies. This paper proposes novel algorithms that will help non-ML experts to understand the explanation produced by XAI tools. The goal is to give the reader an insight into the importance of combining XAI and PHM. This paper is organized as follows: Section 2 states the problem and proposes a solution, Section 3 describes the different approaches that can achieve explainability, Section 4 demonstrates the proposed approach used for this paper, Section 5 explains the algorithms developed, Section 6 introduces three different case studies in which the proposed approach has been applied, Section 7 discusses the solution proposed and finally section 8 summarises conclusions and draws directions for future work.

2. PROBLEM STATEMENT

Many AI and ML diagnostic and predictive applications use black-box type models because their outputs provide better performance than simpler, and therefore more transparent, white-box approaches. However, stakeholders in regulated industries such as Nuclear, base their operational decisions on understanding how models generate their predictions (Preece, Harborne, Braines, Tomsett, & Chakraborty, 2018). In the case of civil nuclear generation, a fault prognostic model may invoke a decision to take the station offline while the fault is investigated, which in turn will incur maintenance costs and lost generation revenues. XAI helps users understand the underlying structure of black-box machine learning models and how they produce their outputs; hence, boosting user's confidence in these models and encouraging them to use them. Unfortunately, most XAI that are in use produce explanations in a technical format that is not easily understandable to a non-ML expert (Bove, Aigrain, Lesot, Tijus, & Detyniecki, 2022), which in the case of power generation, most operational staff will be. Research shows that experts in the application domain tend to trust machine learning models when they are provided with human-friendly explanations that will enable them to understand the rationale of ML models (Bove et al., 2022). Also, there is a requirement for distinctly different explanations for stakeholders in different application domains (Mohseni, Zarei, & Ragan, 2018). To pursue this challenge, this paper proposes a novel application of a set of algorithms that translates explanations generated from XAI tools into human understandable text-based explanations.

3. DIFFERENT APPROACHES TO XAI

Explainability (Interpretability) (Carvalho, Pereira, & Cardoso, 2019) can be achieved through different approaches and they can be classified according to different criteria. In this section, we will explore some of the well-known classifications (Barredo Arrieta et al., 2020):

3.1. Pre-model, During Model and Post-Model

Explainability can be achieved through different complementary approaches. One of these approaches depends on when XAI techniques are applied. They can be applied as: 'Pre-Model', 'During model' and 'Post-model methods' (Stiglic et al., 2020) (Carvalho et al., 2019). 'Pre-model' is done in the first stage of model development after obtaining the related data and before selecting the desired ML model appropriate for the problem statement. The primary goal of using pre-model methods is to understand and describe the data used in ML model and how the data health and structure influence the model. 'During Model' is an approach to ensure explainability through the use of transparent models, which are models that are inherently understandable for humans.(Doran, Schulz, & Besold, 2017) Using transparent models is one approach to achieve interpretability. In these models, humans can easily understand how inputs are mathematically mapped to outputs by having technical knowledge of the model itself and the algorithms used in the models (Molnar, 2020). One drawback of this approach that it is model-specific, and the model design process is limited by the number of representative models available to choose from. Interpretable models include linear regression, logistic regression, generalized linear models, and decision trees (Molnar, 2020). Finally, 'Post model' or 'Post-hoc Methods' is an approach applied after choosing the ML model and after obtaining predictions from these models. Currently, most black-box models are explained using a post-hoc approach. This approach is used for complex models in which humans cannot understand the underlying decision-making mechanism. The advantage of post-hoc approaches is that they do not affect the performance of a complex model as it treats the model as a black-box (Dosilovic, Bri, & Hlupic, 2018). Post-hoc approaches can be primarily classified into three groups:

1. Gradient based attribution methods such as saliency maps (Simonyan, Vedaldi, & Zisserman, 2014) which assign importance scores to each input feature and show which parts of the input are most important.

2. Surrogate Models such as MUSE (Model Understanding through Subspace Explanations) (Lakkaraju, Kamar, Caruana, & Leskovec, 2019). In this approach, black-box models' behavior is explained in sub-spaces defined by specific features that are of user interest.

3. Post-hoc approaches via perturbation: This approach uses perturbations of the input data to generate pairs of inputs and outputs, then uses simple models e.g., linear models to explain the prediction obtained. Examples of techniques that use this approach are LIME and SHAP tools. Shapley Additive exPlanations (SHAP) tool computes feature importance by computing the contribution of each feature to the output obtained. These contributions are calculated using coalitional game theory, where features represent players in a coalition (Molnar, 2020). SHAP tool increases transparency by producing SHAP values for each instance in the data set (Molnar, 2020). SHAP values can be aggregated to provide global interpretability for machine learning models. It is considered an optimal approach for providing interpretability since it is built on a solid theory (Lundberg & Lee, 2017). Some advantages of SHAP is that it is based on a solid theoretical theory and it can provide local and global explainability by providing SHAP force plot for local explainability and SHAP summary plot for global explainability as shown in figure (1) (Lundberg & Lee, 2017). Due to the benefits and wider adoption of SHAP by the AI community, application and development of SHAP values will be a focus of this paper.

3.2. Global and Local explainability

A second approach to classifying interpretability methods is according to the scope of how they assess the underlying model, i.e., from a global or local perspective (hui Li et al., 2022) (Bhatt et al., 2020) (Angelov, Soares, Jiang, Arnold, & Atkinson, 2021). Global interpretability: Global methods help users to logically understand the relationship between all input variables and the predicted output. They help in forming an overall understanding of the behavior of the modal (Doran et al., 2017). Users are able to understand all the different possible outcomes. In contrast, local interpretability provides an explanation for one instance or region of the modeled space, and the associated contribution of that instance or space to the overall output (Bhatt et al., 2020).

3.3. Model-Agnostic and Model-Specific explainability:

In model-agnostic techniques, there is the flexibility to choose any machine learning approach. The machine learning model in a model agnostic approach is treated as a black-box by separating the explanations from the model, thus giving the flexibility to choose any ML model, alongside any representation and explanation (Molnar, 2020) (Angelov et al., 2021). One disadvantage of using model-agnostic techniques is the possibility of having inconsistent local explanations (Ribeiro, Singh, & Guestrin, 2016) (Ribeiro et al., 2016). In model specific techniques, choice is limited to specific models because methods are based on the internal workings of specific models, and it is hence difficult to change to another model (Molnar, 2020).

4. New proposed approach to XAI

The goal is to develop various options for extracting explainability (interpretability) from predictive or diagnostic analytic tools. These extracted explanations being required to be presented to decision-making stakeholders who are non-ML experts in human-friendly context. To achieve this goal, four complementary explanatory stages have been identified (see figure (2)). Each stage is presented next with more details:

1. Data pre-processing: The first stage eases the understanding of the data set used and recognizes the features contained therein. The quality of data is assessed and transformed into an understandable format that can be used later in ML/analytic models.
2. Prediction Models: The second stage is to choose appropriate machine learning prediction model(s). Most ML models are considered black-box models, in which we cannot understand how these models work and how inputs are mapped into outputs. Therefore, there is a need to develop and deploy XAI techniques that generate explanations on predictions made, enabling industry stakeholders to understand the machine learning models adopted.
3. Applying XAI tools: Applying XAI tools to provide understandability of how ML models work and why they produce these predictions. For this paper, a widely adopted post-hoc XAI tool is used. SHAP is applied to provide local and global explanations. SHAP generates more reliable explanations than other XAI tools, and it can provide local explanations for a single predication (e.g. why a specific prediction has been made , what are the most important features contributing to this prediction, and the impact of each feature on the prediction) and a global explanation to provide a holistic understanding of how the ML model works. However, SHAP produces these explanations in the form of complex plots which are not easy to understand, especially for a non-ML expert. This is the rationale for introducing a final stage to translate these explanations into a more understandable format.
4. Generating human understandable explanations: How to communicate explanations to non-ML-experts in the application domain is important. SHAP plots are not always easy to understand, even for a data scientist. This fact leads to the need to translate these plots into a human-understandable context that will result in bridging the gap between ML experts and stakeholders. In this stage, a set of novel algorithms have been developed to translate SHAP local and global explanations to generate human understandable text-based explanations.

5. Automated human-understandable-text generation algorithms

In this section, the algorithms used to translate complex SHAP plots to human-understandable text based explanations are described.

5.1. Translating SHAP local explanation plots

After applying the SHAP tool in order to provide interpretability, the resulting explanations are produced in the form of complex plots. This paper demonstrates a novel approach, where these plots are translated into text-based explanations.

For SHAP local interpretability, a SHAP force plot is produced for a single prediction, providing the most important features, and the impact of each feature on the output at a local level. If a feature has a positive SHAP value, this indicates that the feature value has a positive impact on the prediction. However, if it has a negative value, this indicates that the feature has a negative impact on the prediction and finally, if the SHAP value equals zero, then this feature has no impact on the output.

How this information is translated into human-understandable text is demonstrated in Figure (3), where a flow chart explains how the logic behind the code that has

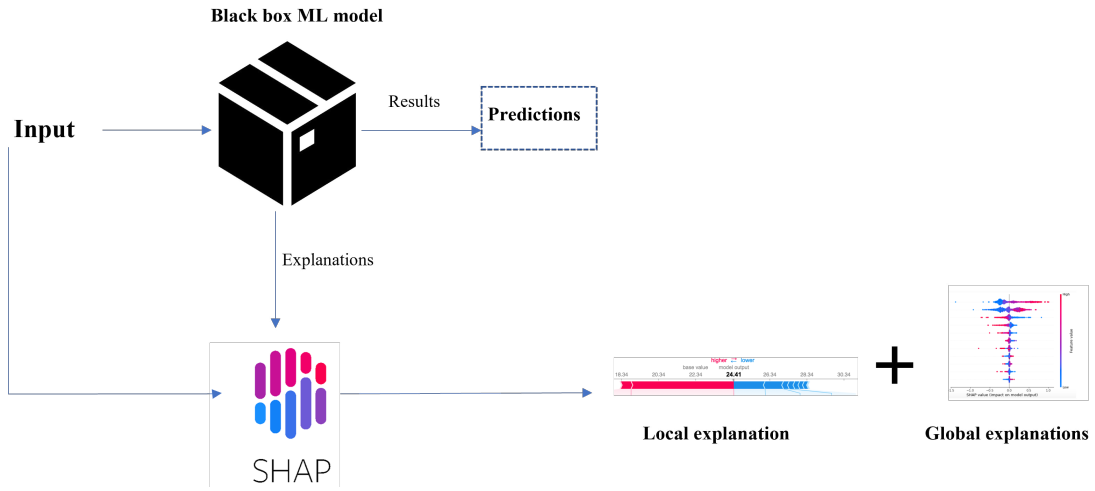

Figure 1. SHAP local and global explanations.

been developed for the automated text-generating process.

5.2. Translating SHAP global explainability plot

For Global explainability, each feature has its SHAP values. The SHAP values for each feature are aggregated and then compared. After that, all the features are ordered according to their aggregated SHAP values. Features with higher aggregated SHAP values have greater impact on the output and vice versa. How this information is translated into human understandable text is demonstrated in Figure (4), where a flow chart explains how the logic behind the code that has been produced for the automated text-generating process.

6. CASE STUDIES

The novel approach to human-understandable XAI, described in Sections 4 and 5, has been applied to three different case studies, as follows.

6.1. Case study 1 : Combined Cycle Gas Turbine (CCGT)

In the first case study, a publicly available data set consisting of operational measurements from a Combined Cycle Gas Turbine (CCGT) generator has been used. An open-source data set has been chosen as the first case study to facilitate easy application of XAI tools and also to make the work reproducible. The CCGT data set was curated over 6 years (2006-2011) and has been previously used to show machine learning models for predicting power output based on environmental conditions (Wood, 2020) (Tüfekci, 2014)

6.1.1. Data pre-processing

The data set composes the following operational measurements from the turbine, generator, and control valves:

1. Ambient pressure (AP).
2. Exhaust Vacuum (V).
3. Ambient temperature (AT).
4. Relative humidity (RH).

These parameters are used to predict the net hourly electrical energy output (PE) of the plant. In this case study, it was shown that the relationship between environmental conditions and power output could be clearly identified and explained. In figure (5), some statistical properties about CCGT data set are provided.

6.1.2. Modelling

Three different candidate models have been implemented for the explainability case-study: linear regression, random forest and XGboost. These ML models were used to predict the output power and their performances were compared using three performance metrics that are usually used to compare performance between different regression models: Root mean squared error (RMSE) metric, which measures the average error performed by the model, R2 score which specifies how close the calculated values are plotted to the actual data values and Mean squared error (MSE). The metrics for Gradient Boosting Regressor showed improvements over the Linear Regression Model and the random regression model. There are 3 key performance metrics (See table 1) used to assess how well each model is performing. After evaluating all the models, XGBoost Regression Algorithm was found to give the best performance with R-squared = 0.97 and RMSE = 3.069.

6.1.3. XAI Application

In this stage, SHAP has been adopted to provide explanations to the ML predictions. It generates explanations in a form of visualizations that are quite complex and are not always intuitive. In figure (6), SHAP summary plot produced using the

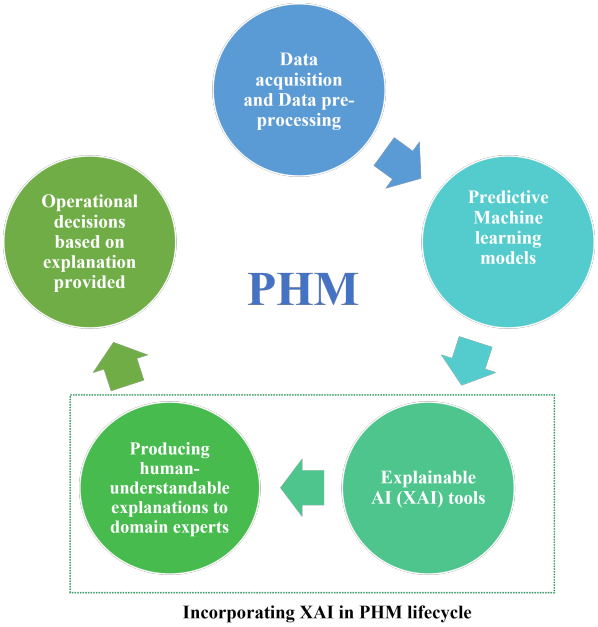

Figure 2. Proposed approach.

Table 1. Comparison of ML models performances.

ML models	R2 Score	MSE	RMSE
Linear regression	0.92	20.637	4.543
Random forest	0.94	15.278	3.909
XGBoost	0.97	9.419	3.069

XGBOOST model. According to SHAP summary plot shown in figure (6), the most important features in descending order are : Temp, Vacuum, Pressure and finally Humidity. The impact of each feature is also shown (e.g. high values (shown in red color) of Temp has a negative impact on the output power causing the output power to decrease while low values(shown in blue) of Temp has positive impact on the output power causing the output power to increase). In the SHAP force plot (SHAP local explainability plot), shown in figure (7), features like Temp, vacuum and pressure (shown in red) causing an increase in the predicted output power. The visual size for each feature in SHAP force plot (size of the arrow) shows the magnitude of each feature's impact. According to this local explanation the most important features in descending order are Temp, Vacuum, Pressure and finally Humidity.

6.1.4. Generation of human-understandable explanations

As described in Section 5, a set of algorithms have been developed to achieve the task of translating SHAP plots into text-based explanations for ease of comprehension. In figure (8), an example of the text-based explanations generated by translating the SHAP local explainability plot (SHAP force plot) is shown in figure (7). The text-based explanation clearly describes the most important features of the plot and the associated impact from each feature value contributing towards the predicted output power (e.g. Temp =11.37 is considered a low value after comparing it to the mean value of Temperature in the data set and has a positive impact on the output power pushing the output power value higher). Figure(9) shows the automated text-based explanations for the SHAP summary plot shown in figure(6). While Figure (10) shows summary statistics for each feature, including: the number of values for each feature that have no impact on the output power; the number of values that are considered high and have high/low positive impact on the output pushing the output power to increase; and the number of values for each feature that are considered low and have a high/low negative impact on the output power, causing the prediction to decrease.

6.2. Case study 2: Boiler Feed Pump Gearbox Data set

A gearbox is a mechanical device used to increase or decrease the speed of another part connected to it along a rotating drive-train. The objective of this case study is to apply XAI tools to provide explanations to predictive models applied to a gearbox data set related to a boiler feed pump and then ease the ability to understand these explanations through the application of the auto-generated novel text-based algorithms proposed in this paper. The modelling aim was to investigate how the controlling stop-valve position values affect rms-vibration and the operational consequences associated with increased stop-valve position. Increased vibration in boiler feed pump lead to decrease in the performance of

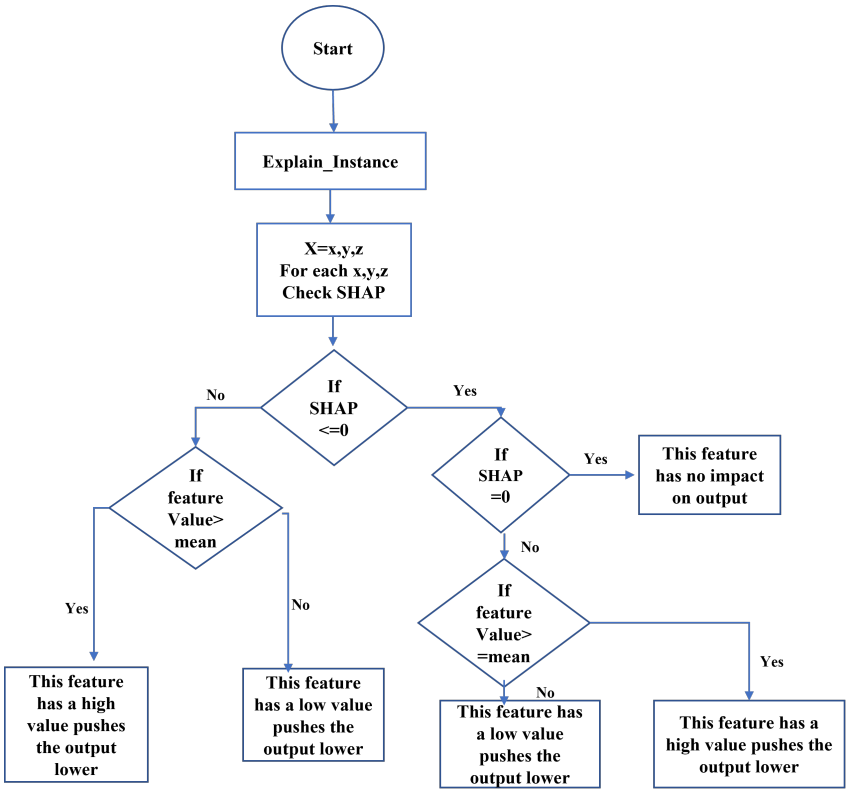

Figure 3. Flowchart to translate SHAP force plot.

pump and result in damage to some pump parts. In this case study, it was shown that the relationship between stop-valve position and rms-vibration could be clearly identified and explained.

6.2.1. Data pre-processing

The data set used compromises of different valve positions and rms-vibration to investigate if there is a correlation between the valve position and rms-Vibration. This data was collated and provided by a real operational boiler feed pump in the power generation industry. The following operational measurements were used to create the predictive model:

1. Stop-valve position.
2. Rms-vibration.

Before machine learning prediction models can be used, the time-series data set has been re-framed as a supervised learning problem, resulting in a sequence of input and output pairs. Reframing the data set removes the complexities around the prediction problem and can give more reliable forecasts. After re-framing the data set to a supervised learning problem, the following operational measurements used to predict rms-vibration(t+1) are: stop-valve-position(t-2), rms-vibration(t-2), stop-valve-position(t-1), rms-vibration(t-1), stop-valve-position(t), rms-vibration(t) and stop-valve-position(t+1).

Table 2. Comparison of ML models performances.

ML models	R2 Score	MSE	RMSE
Linear regression	0.96	0.001417	0.03766
Random forest	0.96	0.00144	0.0379
XGBoost	0.95	0.00151	0.03889
Ensemble Model	0.96	0.001455	0.03815

6.2.2. Modelling

Similar to the previous case study, the same three different ML models were assessed for their effectiveness: Linear regression, Random Forest, and XGBoost - all being used to predict rms-vibration(t+1). These ML models were then combined using an averaging ensemble model to improve the overall performance. Performances have been compared as seen in table (2) using three different performance metrics. As shown in table (2), Ensemble model has not improved the overall performance and linear regression has the best performance of the all models. It is concluded from these results that a linear regression model should be selected to create the SHAP values.

6.2.3. XAI application

This case study adopted SHAP to provide explanations for the ML predictions. Figure (11) is the SHAP summary plot

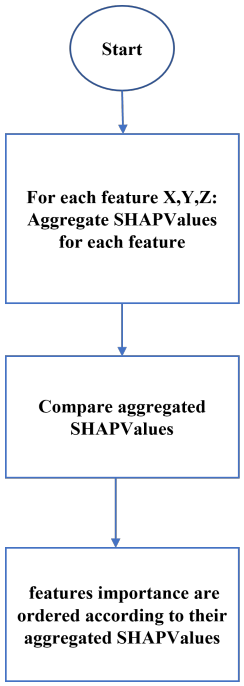

Figure 4. Flowchart to translate SHAP Summary plot.

	Temp	Vacuum	Pressure	Humidity	Power
count	9568.000000	9568.000000	9568.000000	9568.000000	9568.000000
mean	19.651231	54.305804	1013.259078	73.308978	454.365009
std	7.452473	12.707893	5.938784	14.600269	17.066995
min	1.810000	25.360000	992.890000	25.560000	420.260000
25%	13.510000	41.740000	1009.100000	63.327500	439.750000
50%	20.345000	52.080000	1012.940000	74.975000	451.550000
75%	25.720000	66.540000	1017.260000	84.830000	468.430000
max	37.110000	81.560000	1033.300000	100.160000	495.760000

Figure 5. Statistical details about CCGT data set.

using linear regression. The most important features are shown, with rms-vibration(t) being the most important feature and rms-vibration(t-2) the least important. The SHAP summary plot also shows the correlation between each feature and the output (e.g. rms-vibration(t) is positively correlated with the output, the higher the rms-vibration(t) is, the higher the output and similarly for rms-vibration(t-1)). In the SHAP force plot (local explainability plot), shown in figure (12), the most important feature for this prediction is shown in red: rms-vibration(t), having a positive correlation with the output (causing the output to increase). On the other hand, rms-vibration(t-1) shown in blue has a negative correlation with the output, causing the output to decrease.

6.2.4. Generation of human-understandable explanations

The techniques from Section 5 were then used to generate automated-text-based explanations that are easy to understand. In figure (13), an example of the text-based explanation generated corresponding to the SHAP force plot (local explanations plot) produced in figure (12). In figure (13), the most important features affecting the output for a specific instance are listed. Also, the impact of each feature value and whether this feature value pushes the output value higher/lower is shown (i.e. rms-vibration(t) has a low value for this instance that pushes the output higher). In Figure (14), a text-based explanation corresponding to SHAP summary plot shown in figure (11) is provided, denoting the most important features globally for the prediction model. Figure (15) shows summary statistics for some of the features, including: the number of values for each feature that have no impact on the output, number of values that are considered high and have high/low positive impact on the output pushing the output to increase and the number of values for each feature that are considered low and have a high/low negative impact on the output causing the prediction to decrease.

6.3. Case study 3: Thrust bearing wear predictive model

In this case study real condition monitoring data from feedwater pumps has been used to anticipate thrust bear wear (denoted ”median-TB”) given operating parameters such as flow (”mean-Flow”) and head (”mean-Head”).The data set used compromises of different values of flow and head. The data set is used to predict thrust bearing wear.

6.3.1. Data pre-processing

The data set used compromises of different values of flow and head. This data set is used to predict thrust bearing wear. Similar to the pre-processing stage for case-study two described in section 6.2 the time series data has been re-framed to a supervised learning problem from a sequence to pairs of input and output sequences. The following operational measurements are used to predict median-TB(t+1): mean-Flow(t-2), mean-Head(t-2), median-TB(t-2), mean-Flow(t-1), mean-Head(t-1), median-TB(t-1), mean-Flow(t), mean-Head(t), median-TB(t), mean-Flow(t+1) and mean-Head(t+1).

6.3.2. Modelling

Similar to the previous case-studies, the same three ML models are assessed for their predictive accuracy: linear regression, Random Forest, XGBoost have been used to predict thrust bearing wear (median-TB(t+1)). Then, ML models have been combined using the same averaging ensemble model to investigate whether or not the overall performance will be improved which in this case it didn’t . Performances

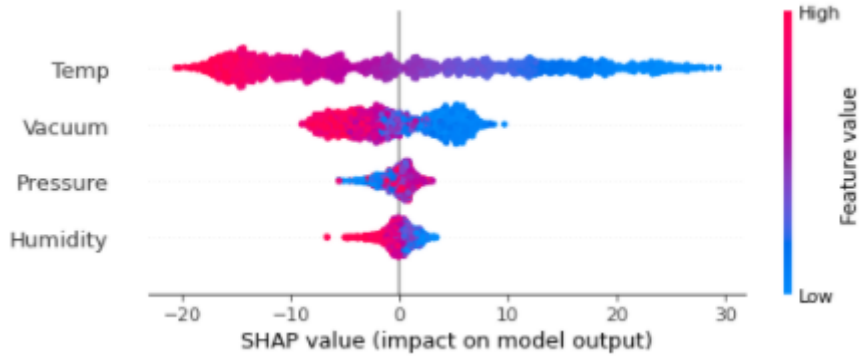

Figure 6. SHAP summary plot for case study 1.

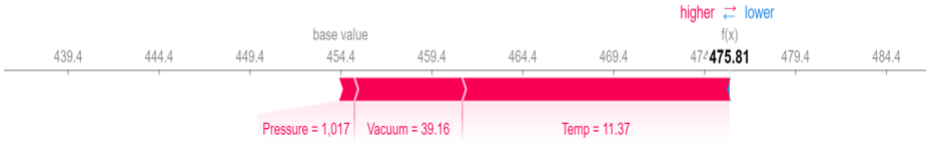

Figure 7. SHAP force plot for case study 1.

```
features affecting the output power in order are
  Temp
  Vacuum
  Pressure
  Humidity
impact of each feature value on output
Temp    11.37 Low value pushes the output power higher
Vacuum    39.16 Low value pushes the output power higher
Pressure    1016.54 High value pushes the output power higher
Humidity    87.05 High value pushes the output power lower
```

Figure 8. Text-based explanations for SHAP force plot in case-study 1

```
Most influential features are
Temp
Vacuum
Pressure
Humidity
```

Figure 9. Text-based explanations for SHAP summary plot in case-study 1

```
Global interpretability for  Temp
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 45.03657262277952
% of points that have high values and positive impact on output = % 0.41797283176593525
% of points that have low values and Negative impact on output = % 1.4629049111807733
% of points that have high values and Negative impact on output = % 53.08254963427377
Global interpretability for  Vacuum
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 44.40961337513062
% of points that have high values and positive impact on output = % 0.8881922675026124
% of points that have low values and Negative impact on output = % 6.948798328108673
% of points that have high values and Negative impact on output = % 47.7533960292581
Global interpretability for  Pressure
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 21.26436781609195
% of points that have high values and positive impact on output = % 26.854754440961337
% of points that have low values and Negative impact on output = % 31.765935214211076
% of points that have high values and Negative impact on output = % 20.114942528735632
Global interpretability for  Humidity
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 43.155694879832815
% of points that have high values and positive impact on output = % 9.770114942528735
% of points that have low values and Negative impact on output = % 2.507836990595611
% of points that have high values and Negative impact on output = % 44.56635318704284
```

Figure 10. Text-based explanations representing simple statistics for SHAP summary plot in case-study 1

are compared as shown in table (3). Linear regression model has the best performance with the least mean squared error (MSE).

6.3.3. XAI application

Applying SHAP techniques to provide explanations to machine learning predictions produced the following results. Figure (16) depicts the SHAP summary plot, showing the most important features contributing to model predictions. The plot shows SHAP values for each feature and the impact these features have on the model predictions. The most important features for this model from the global explanation perspective in descending order are: mean-Flow(t+1), median-TB(t), mean-Head(t+1), ..., and lastly mean-Head(t-1) as depicted in figure (16). From the SHAP summary plot mean-Flow(t+1) is positively correlated to the output, the

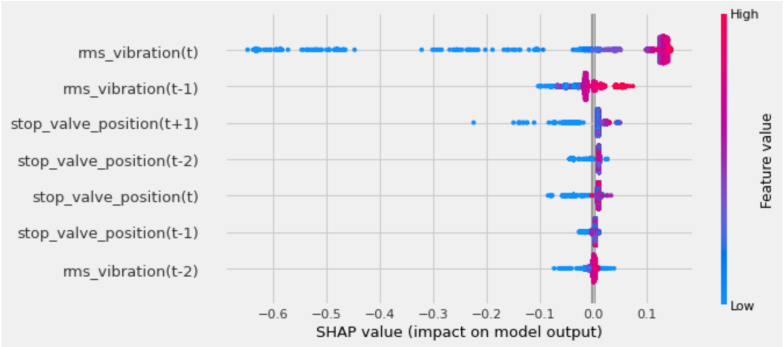

Figure 11. SHAP summary plot for case study 2

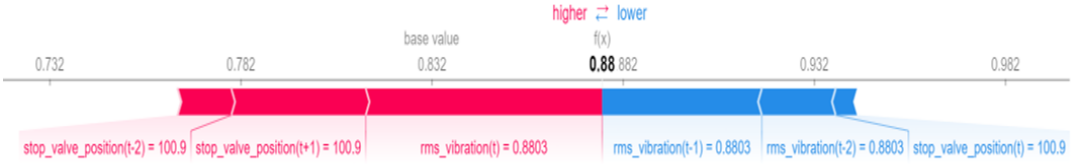

Figure 12. SHAP force plot for case study 2

Table 3. Comparison of ML models performances.

ML models	R2 Score	MSE	RMSE
Linear regression	0.9978	0.0000035	0.002
Random forest	0.997	0.0000042	0.00207
XGBoost	0.997	0.0000045	0.0021
Ensemble model	0.997	0.0000041	0.0021

higher the mean-Flow(t+1), the higher the output. Figure (17) shows the local explanation generated by SHAP for a single prediction using the linear regression model. In the SHAP force plot figure (17), the most important feature for this prediction (shown in red) is mean-Flow(t+1), having a positive correlation with the output (causing the output to increase). The second most important feature is median-TB(t) (shown in red), having a positive correlation with the output (causing the output to increase for this prediction).

6.3.4. Generation of human-understandable explanations

The techniques from Section 5 were used to generate automated-text-based explanations that are easy to understand. In figure (18), an example of text-based generated corresponding to SHAP local explanations plot produced above and shown in figure (17). In Figure (19), text-based explanations corresponding to the SHAP summary plot shown in figure (16), showing the most important features globally for the prediction model.

7. Discussion

The aim of this work is to introduce XAI techniques into PHM systems. In this paper, a new approach has been proposed to produce a human-understandable format of SHAP produced explanations. Compared to other related literature which lacks human understandability, this approach makes it easier for non-ML experts to understand the results from explainability tools. The authors propose that the text-based representation of the SHAP process is easier and more intuitive to interpret because they allow non-ML experts to understand and engage with how ML models work. These text-based explanations will enable stakeholders to understand the impact of each input and the operational consequences associated with different inputs/values. The proposed approach has been used in three different case studies and demonstrates the provision of a human-friendly form of explanations to non-ML experts.

8. Conclusions and Future Work

Exploiting the application of XAI tools in PHM can lead to increased confidence in PHM systems, encourage their adoption, and ultimately meet the assurances and quality required for PHM system deployment in safety-critical industries such as nuclear. In this paper, through the development and demonstration of a novel approach to the interpretation of a well-known post-hoc XAI technique (SHAP), it has been shown that explanations in a 'human-friendly' format can aid stakeholders (who are not necessarily ML experts) to rapidly interpret the technical explanations provided

```
  instance no 0  taken at time:  2014-09-20 17:10:00
Turbine speed at this time:
 TurbinespeedA=  4517.489258
 TurbinespeedB=  4521.881348
features affecting the rms-vibration in order are
  rms_vibration(t)
  rms_vibration(t-1)
  stop_valve_position(t+1)
  rms_vibration(t-2)
  stop_valve_position(t-2)
  stop_valve_position(t)
  stop_valve_position(t-1)
impact of each feature value on output
stop_valve_position(t-2)   100.8535309 High value pushes rms-vibration higher
rms_vibration(t-2)   0.8802593 Low value pushes rms-vibration lower
stop_valve_position(t-1)   100.8539429 High value pushes rms-vibration higher
rms_vibration(t-1)   0.8802593 Low value pushes rms-vibration lower
stop_valve_position(t)   100.8543549 High value pushes rms-vibration lower
rms_vibration(t)   0.8802593 Low value pushes rms-vibration higher
stop_valve_position(t+1)   100.8547668 High value pushes rms-vibration higher
```

Figure 13. Text-based explanation for SHAP force plot for case study 2.

```
Most influential features are
rms_vibration(t)
rms_vibration(t-1)
stop_valve_position(t+1)
stop_valve_position(t-2)
stop_valve_position(t)
stop_valve_position(t-1)
```

Figure 14. Text-based explanations for SHAP summary plot for case study 2

```
Global interpretability for  rms_vibration(t-2)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.8630648997621474
% of points that have high values and positive impact on output = % 81.97757390417941
% of points that have low values and Negative impact on output = % 16.71083927964662
% of points that have high values and Negative impact on output = % 0.44852191641182465
Global interpretability for  stop_valve_position(t-1)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 3.9755351681957185
% of points that have high values and positive impact on output = % 0.12232415902140673
% of points that have low values and Negative impact on output = % 0.0
% of points that have high values and Negative impact on output = % 95.90214067278288
Global interpretability for  rms_vibration(t-1)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.12232415902140673
% of points that have high values and positive impact on output = % 81.14169215086646
% of points that have low values and Negative impact on output = % 17.44478423377506
% of points that have high values and Negative impact on output = % 1.291199456337071
Global interpretability for  stop_valve_position(t)
No of points that have no impact on output= % 0.0
 % of points that have low values and positive impact on output = % 0.013591573224600747
% of points that have high values and positive impact on output = % 83.7037037037037
% of points that have low values and Negative impact on output = % 3.961943594971118
% of points that have high values and Negative impact on output = % 12.320761128100578
Global interpretability for  rms_vibration(t)
```

Figure 15. Text-based explanations for some of the features in case study 2 .

by black-box ML models that may comprise a PHM methodology. This subsequently increases their confidence in adopting the model that may have produced new prognostic insight for operational decisions. This new approach is intended to support end-users (who are not ML experts) interpret the outputs from SHAP and this benefit is realised through the creation of new algorithms that auto-generate text-based explanations based on SHAP summary and force plot outputs. These text-based explanations provide a more intuitive means of interpreting SHAP outputs, which are more generally intended for data scientists or other practitioners familiar with the field of study. The approach developed has been applied to three case-studies – two (2 and 3) of which are based upon operational data from a nuclear power station. They demonstrate that it is possible to produce more intuitive explanations than the standard graphical outputs produced by SHAP tools. These more intuitive text-based explanations can henceforth be more easily understood by the end-user of the related PHM algorithms, who may be unfamiliar with both: the ML predictive algorithm in its own right but also the methodology and format associated with SHAP. In addition, during the investigation associated with this paper, the authors have identified some limitations of the proposed approach that can be further improved to produce more robust and reliable explanations, and which are the focus on on-going work. One limitation identified, and associated with using a correlating post-hoc tool such as SHAP, is the absence of the ability to causally link the correlations identified by SHAP to related physical phenomena. Introducing causality into post-hoc XAI tools will help in providing more reliable explanations, both by related the correlations to the underpinning physics but also by potentially providing explanations in specific engineering domain contexts. In addition to these causality investigations, the authors have a further aim to develop additional/improved means of intuitively representing and subsequently interrogating AI explanations. Building on the content of the work described in this paper, the authors are currently developing techniques to auto-generate graph-based representations of the semantic knowledge embedded within AI explanations. The intended methodology aims to continue improving on

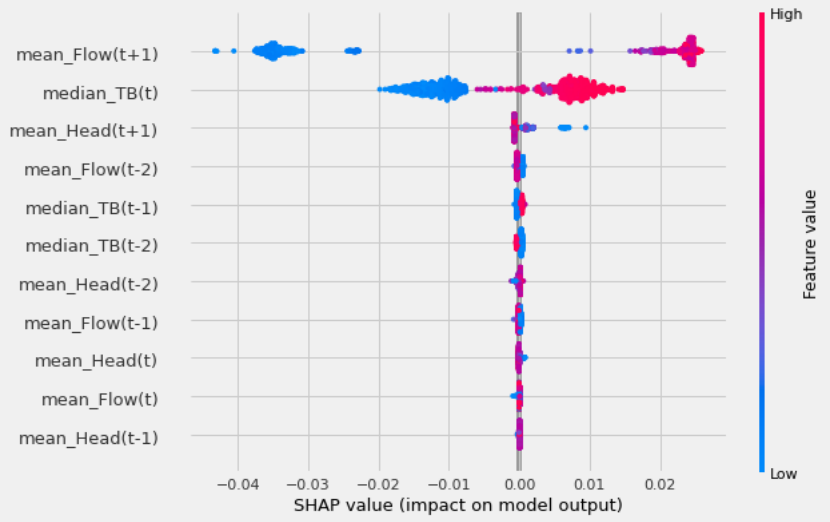

Figure 16. SHAP summary plot for case study 3.

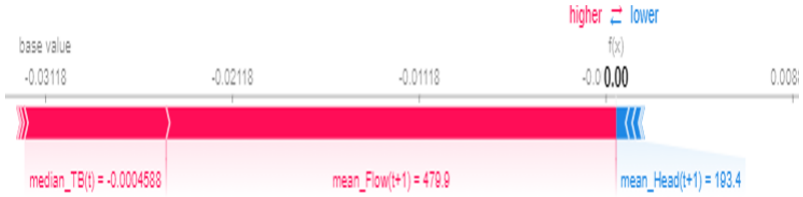

Figure 17. SHAP force plot for case study 3.

```
  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 18. Text-based explanation for shap force plot for case study 3

```
Most influential features are
mean_Flow(t+1)
median_TB(t)
mean_Head(t+1)
mean_Flow(t-2)
median_TB(t-1)
median_TB(t-2)
mean_Head(t-2)
mean_Flow(t-1)
mean_Head(t)
mean_Flow(t)
```

Figure 19. Text-based explanation for shap summary plot for case study 3

how (non-ML expert) end-users can adopt PHM through explanation of the related AI technique but in parallel also facilitate machine interactions and interfacing with the software-based explanation process. It is proposed that providing a means of machine interface to the explanation process can lead to the inclusion of techniques such as query language and more sophisticated graph manipulation; ultimately resulting in more insight and knowledge discovery, both for nuclear engineers hoping to adopt ML-based PHM techniques but also more generally in industries with a similar deficit of ML capability.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support received from the UK's National Physical laboratory (NPL) and from the Advanced nuclear research centre (ANRC) at the University of Strathclyde, Glasgow.

REFERENCES

Angelov, P., Soares, E., Jiang, R., Arnold, N., & Atkinson, P. (2021, 09). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*. doi: 10.1002/widm.1424

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, *58*, 82-115. doi: https://doi.org/10.1016/j.inffus.2019.12.012

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (p. 648–657). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3351095.3375624

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2022). Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *27th international conference on intelligent user interfaces* (p. 807–819). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3490099.3511139

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). doi: 10.3390/electronics8080832

Doran, D., Schulz, S., & Besold, T. R. (2017). *What does explainable ai really mean? a new conceptualization of perspectives.* arXiv. doi: 10.48550/ARXIV.1710.00794

Dosilovic, F. K., Bri, M., & Hlupic, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210-0215.

hui Li, X., Cao, C. C., Shi, Y., Bai, W., Gao, H., Qiu, L., ... Chen, L. (2022). A survey of data-driven and knowledge-aware explainable ai. *IEEE Transactions on Knowledge and Data Engineering*, *34*, 29-49.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (p. 131–138). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3306618.3314229

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A survey of evaluation methods and measures for interpretable machine learning. *ArXiv*, *abs/1811.11839*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Preece, A. D., Harborne, D., Braines, D., Tomsett, R. J., & Chakraborty, S. (2018). Stakeholders in explainable ai. *ArXiv*, *abs/1810.00184*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-agnostic interpretability of machine learning.* arXiv. doi: 10.48550/ARXIV.1606.05386

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, *abs/1312.6034*.

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning based prediction models in healthcare. *CoRR*, *abs/2002.08596*.

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power Energy Systems*, *60*, 126-140. doi: https://doi.org/10.1016/j.ijepes.2014.02.027

Wood, D. (2020, 03). Combined cycle gas turbine power output prediction and data mining with optimized data matching algorithm. *SN Applied Sciences*, *2*. doi: 10.1007/s42452-020-2249-7