O. Serradilla, E. Zugasti, C. Cernuda, A. Aranburu, J. R. de Okariz and U. Zurutuza, "Interpreting Remaining Useful Life estimations combining Explainable Artificial Intelligence and domain knowledge in industrial machinery," 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2020, pp. 1-8.

DOI: https://doi.org/10.1109/FUZZ48607.2020.9177537

Interpreting Remaining Useful Life estimations combining Explainable Artificial Intelligence and domain knowledge in industrial machinery

Oscar Serradilla
Electronics and Computing Department
Mondragon Unibertsitatea
Mondragon, Spain
oserradilla@mondragon.edu

Ekhi Zugasti
Electronics and Computing Department
Mondragon Unibertsitatea
Mondragon, Spain
ezugasti@mondragon.edu

Carlos Cernuda
Electronics and Computing Department
Mondragon Unibertsitatea
Mondragon, Spain
ccernuda@mondragon.edu

Andoitz Aranburu
Innovation in Automotive Division
Fagor Arrasate
Mondragon, Spain
a.aranburu@fagorarrasate.com

Julian Ramirez de Okariz
Mechanical Engineering
Koniker
Mondragon, Spain
j.rokariz@koniker.coop

Urko Zurutuza
Electronics and Computing Department
Mondragon Unibertsitatea
Mondragon, Spain
uzurutuza@mondragon.edu

***Abstract*—This paper presents the implementation and explanations of a remaining life estimator model based on machine learning, applied to industrial data. Concretely, the model has been applied to a bushings testbed, where fatigue life tests are performed to find more suitable bushing characteristics. Different regressors have been compared Environmental and Operational Condition and setting variables as input data to prognosticate the remaining life on each observation during fatigue tests, where final model is a Random Forest was chosen given its accuracy and explainability potential. The model creation, optimisation and interpretation has been guided combining eXplainable Artificial Intelligence with domain knowledge.**

Precisely, ELI5 and LIME explainable techniques have been used to perform local and global explanations. These were used to understand the relevance of predictor variables in individual and overall remaining life estimations. The achieved results have been process knowledge gain and expert knowledge validation, assertion of huge potential of data-driven models in industrial processes and highlight the need of collaboration between expert knowledge technicians and eXplainable Artificial Intelligence techniques to understand advanced machine learning models.

***Index Terms*—Explainable Artificial Intelligence, interpret, Machine Learning, data-driven model, Remaining Useful Life, prognosis, industrial process, domain knowledge**

I. Introduction

Nowadays, we are in the fourth revolution denominated as Industry 4.0 (I4.0), which is based on Cyber Physical Systems (CPS) and Industrial Internet of Things (IIoT). Its objective is to improve and optimise industrial processes by adapting software, sensors and intelligent control units to meet their requirements [1].

One of the main opportunities identified in I4.0 is the maintenance optimisation by applying data-driven techniques to the massive amount of process data. This enables predictive and proactive maintenance strategies, knowledge discovery and process optimisation [2]. There are three type of approaches to address the aforementioned challenges given their underlying technique: expert-knowledge, data-driven and hybrid.

During the last years, Artificial Intelligence (AI) and Machine Learning (ML) models usage in industry has increased due to the increase of available data and the difficulty of modelling industrial data and machine behaviour relying only on expert knowledge based models, due to their variability.

Many data-driven publications in industry nowadays are based on complex ML models given their high accuracy. Conversely, according to Zadeh's Principle of Incompatibility [3]: *as the complexity of a system increases, our ability to make precise and yet significant statements about its behavior diminishes until a threshold is reached, beyond which precision and significance (or relevance) become almost mutually exclusive characteristics*, what makes these ML models difficult to interpret.

The development of data-driven models in industry faces several challenges non-existent in other domains. One main challenge is the variability of machine data: real machine behaviour varies from theoretical knowledge due to tolerances, mount adjustments, variations in Environmental and Operational Conditions (EOC) and other factors. This makes machines built under the same specifications behave differently.

To address the stated issues, this paper presents a work where XAI techniques have been used to guide a prognosis ML model creation and interpretation in an industrial use-case of bushing testbeds. First, background and related work are briefly presented in Section II. Section III defines the conducted research work, its limitations and development steps. Section IV describes the use-case and dataset used to

This research was sponsored by Fagor Arrasate S.Coop in collaboration with the Business R&D Unit Koniker S.Coop.

perform the experiments of this work. Section V presents and discusses the research results. Finally, Section VI presents general conclusions and possible future research lines.

II. Background and related work

A. Industrial models

As stated in introduction, the types of models used in industry can be classified by their methodology [4]:

1) Expert knowledge/model-based: use system's failure mechanisms knowledge to build a mathematical description of its degradation, resulting in a white-box approach that is easy to translate to physical meaning. Conversely, they are difficult to implement in complex systems. The following works present two types of models. Li, Y., T. R. Kurfess, and S. Y. Liang [5] use a stochastic prognostics for Remaining Useful Life (RUL) prediction on rolling element bearings. Oppenheimer, Charles H., and Kenneth A. Loparo [6] propose a physics-based approach for diagnostics and prognostics of cracked rotor shafts.

2) Data-driven: try to predict the machine state based on sensor data. They are composed of statistical methods, reliability functions and artificial intelligence methods. These are composed by grey-box models as fuzzy rule-based systems [7] or bayesian networks [8], and black-box model like Random Forest [8], eXtra Gradient Boosting (XGBoost) or deep learning models. These data-driven approach does not need to understand complex system's physics. Usually, they are more precise than expert-knowledge based on complex systems but their results are difficult to relate to physical meaning. The following two works are based on this type of models. Si et. al in [9] use a Wiener-process based degradation model with a recursive filter algorithm to estimate the RUL. Guo, Liang, et al. in the publication [10] use a health indicator composed by a Recurrent Neural Network (RNN) for RUL prediction in bearings. Pichler et al. [11] present a fault detection approach based on autocorrelation using logistic regression and Support Vector Machines in reciprocating compressor valves.

3) Hybrid: approach combines knowledge-based and data-driven approaches, resulting in a grey-box methodology. Two relevant studies that use hybrid approach are the following. A framework that combines model-based and data-driven methods for RUL prediction on lithium-ion batteries proposed by Liao, Linxia and F. Kottig [4]. An integrated prognostics method composed by a hybrid model that qualifies uncertainty for gear remaining life prediction published by Zhao, Fuqiong, Z. Tian, and Y. Zeng [12].

It is hard to compare performance among works given their performance is tied to used datasets and their characteristics. Therefore, they commonly run different models to compare performance.

B. Explainable Artificial Intelligence

1) Background: A. Adadi and M. Berrada in the work [13] propose a classification of eXplainable Artificial Intelligence (XAI) techniques based on three characteristics that they have:

- **Complexity**: the more complex the model is, the more difficult it is to interpret.
- **Scoop**: where *global interpretability* techniques analyse model's overall logic and reasoning, and *local interpretability* techniques analyse model's decision based on individual data observations.
- **Level of dependency** where 2 types of interpretability techniques are distinguished: *model-specific* take advantage of the particularities of the model handling it as white-box and *model-agnostic* that are general techniques applicable to different models since they treat models as black-box.

The work by Arrieta et al. [14] presents an overview of XAI, gathering concepts and taxonomies, and presenting opportunities and challenges of the field. J. M. Alonso, Ciro Castiello, and Corrado Mencar present a bibliometric analysis paper about XAI works [15], concluding that one third of works belong to the fuzzy logic field. J. M. Alonso, L. Magdalena, and S. Guillaume present a methodology to generate interpretable linguistic knowledge based on fuzzy logic; combining expert knowledge and data-extracted knowledge [16].

2) Common techniques: There are different XAI techniques available, which the book by Samek et al. [17] classifies into four groups regarding their underlying technique. Approximation using local surrogate functions samples neighbours of an observation to perform local explainability as Local Interpretable Model-Agnostic Explanations (LIME) [18] does. Local perturbations analyse how perturbations on an observation change model's output, i.e. Sensitivity Analysis feature importance technique proposed for Random Forests [19], that was generalised afterwards [20]. Propagation-based approaches integrate in model structure using local redistribution rules based on methods as Layer-wise Relevance Propagation (LRP) [21]. Finally, Meta-explanations use techniques like Spectral relevance analysis (SpRAY) [22] and SHapley Additive exPlanations (SHAP) [20] that aggregate local explanations to obtain global explanations.

3) XAI in industry: industrial companies and their technicians need a tool that is accurate but at the same time understandable to trust it and know in which cases it works well and in which it does not.

Nowadays there are very few published works that use explainability to understand complex and black-box data-driven ML models in industry. Most publications lack of XAI techniques since they are based on classifiers to classify new data into already known failure types, which makes the diagnosis step denominated as Root Cause Analysis (RCA) straightforward. However, some works use XAI to explain how these algorithms work. For instance, the publication by J.R. Rehse, N. Mehdiyev and P. Fettke [23] presents an explainability work applied to a smart factory. They implement a DL-based Predictive Maintenance (PdM) system using Long Short Term Memory (LSTM). They focus on post-hoc explanations so that their experts can trust the decisions taken by the model. As authors state, they perform a *binary classification problem,*

using global feature importance and local explanations for process outcome predictions obtained by the applied deep neural network, complemented with textual explanations. They state that they are working on calculating saliency maps, which shows which parts of input data have influence on model output, and visualisation techniques on original or latent space of stacked Autoencoders to make them more interpretable.

Another application of XAI techniques in industry is the one proposed by M. Carletti, C. Masiero, A. Beghi and G.A. Susto [24], a feature importance evaluation approach designed for Isolation Forest to understand the detected anomalies by the model and perform diagnosis in unsupervised way.

The presented works show applications of ML models combined with XAI techniques to meet industrial requirements.

III. Conducted research

This section explains the conducted research work and steps to achieve the results presented in the following section.

A. Work and limitations

This work focuses on the application of a ML model to predict the remaining useful life of fatigue tests based on experiment characteristics and their monitored variables. Concretely, the model predicts the remaining time of experiments in each data observation, considering that components are useful until fatigue failure happens. This model will be used to understand which are the features that have more influence in the experiments' remaining time and find relations among experiments.

Its main contribution is intended to reduce the number of experiments necessary to categorise each tested characteristic, helping to optimise experiment design strategy. This will reduce the time and resources needed to test each experimented characteristics by only focusing on the most relevant variables.

To achieve that, the influence of variables in the duration of experiments will be analysed to infer knowledge. This process will combine expert-knowledge and XAI techniques to explain model's decisions. Furthermore, explainability techniques will be used to analyse experiment characteristics grouped by feature relevance.

B. Development steps

The development steps of this work are based on CRISP-DM [25], adapted to industrial data characteristics and focused on understanding the process, dataset and created model with the objective of explaining its predictions.

The first stage was to understand how the physical process works theoretically and practically, understanding its underlying mechanics. This permits to learn insight about the data to be analysed afterwards, with the objective of learning about the variables that influence the experiments and how they are correlated.

The second stage was to perform a preliminary data analysis of the dataset in order to understand the data and dive into the process combined with domain knowledge.

The third stage was generating the ML-based remaining life estimator. Its development consisted of standard ML development steps that are explained in the following paragraphs.

We performed preprocessing by replacing each Not Available (N/A) value with the mean of that variable in the train dataset, encoded categorical variables and decided to not normalise or standardise the variables because the chosen model did not need it, so the original magnitude and distribution was kept. The dataset partition for model training and tested was performed in the following way: a 10-fold cross-validation stage. It has been also used to analyse not only models prediction stability through data partitions but also analysis whether data random split stability; concretely 80% of experiments for training and remaining 20% for testing.

After preprocessing, feature engineering was performed. Feature selection was done to reduce the number of predictor features because a high dimension of correlated variables adds complexity and information redundancy. Firstly, we removed the variables with zero variance, aka constants. After that, we made a correlation analysis using heatmaps and, together with expert knowledge technicians, we removed derived variables since they were combination of other variables, selecting only the sensor and setting variables. After that, we tested several feature selection techniques to compare them and choose the one that achieved best error score. Concretely, the following techniques have been used and compared. Recursive Feature Elimination, to remove one feature at a time using default model feature importance. Selecting the most relevant feature incrementally using ELI5's [26] black-box feature importance for the model trained with all features, starting with the highest relevant feature, then adding the second one, and so on and so forth until all features are selected. minimum Redundancy Maximum Relevance (mRMR), based on mutual information [27], that maximises the relevance of selected features to the target feature and at the same time minimises the redundancy among them. Selecting the k-best features according to mutual information and ftest indicators. Finally, using Lasso and Ridge linear regressors' feature weights as importance for feature selection.

Feature extraction and dimensionality reduction methods could have also been used to reduce even more the dimensionality and enhance model performance, but these new features are more difficult to understand than original process data. Therefore, we decided not to use this techniques, also supported on the fact that we had already reduced the dimensionality of the selected variables to a small subset and afterwards we saw the model performed correctly with them.

The ML model selected for the work should be a regressor that, given an observation of experiment data predicts the number of seconds remaining until it ends. We decided to compare common ML regression models of State of the Art with the exception of deep learning models, to fit the dataset and use-case requirements. The models analysed in the dataset have been Gaussian Naive Bayes, Linear Support Vector Regressor, K-Neighbors Regressor, Linear Discriminant Analysis, Random Forest Regressor and XGBoost Regressor,

using their default parameters for python's library scikit-learn. All models have been used for supervised regression problems and use different techniques to predict the target class. Despite the last two models are tree-based algorithms, their differences have effect on model explainability. Random Forest Regressor is a *bagging* ensemble method, which underlies on the combination of less precise uncorrelated models to improve the precision and generalisation in an assembled model. It aggregates many random *Regressor Trees* trained with different features and data subsets, forming a *forest*. The regressions are done by averaging trees' regressions. XGBoost Regressor works similarly to Random Forest, ensembling trees to create a forest. Conversely, this is a *boosting* ensemble method, which aggregates new trees to fix the errors generated by existing trees based on gradient boosting method.

After that, we chose the evaluation metrics for the model to measure the overall performance of the model. Then, we run the model, analysed its accuracy ranking using the chosen score, analysed the explainability and iterated in these steps until an acceptable model was obtained.

The explainability of this model was computed and interpreted using local and global explainability methods. Concretely, ELI5's global explanations were used to analyse the global influence of each variable in model predictions together with expert knowledge. This technique was chosen since it removes a variable, concretely it shuffles attribute values to randomise the chosen variable, and analyses model's performance decrease. Local explanations were also calculated using the library LIME, where the contribution of each predictor to model's output is calculated. This is done by measuring how each variable contributes either positively or negatively to the prediction, fitting a linear regression to its neighbour observations. The reason why linear models are used is that these are inherently interpretable, so they are used to analyse the behaviour of small perturbations in the observations.

The fourth and last stage takes as reference the final model developed in the previous step, adapting it to fit the data grouped by experiments of similar characteristics. First, we chose by which experiment characteristics we were interested to group the data based on similarity. Afterwards, one model was created for each group. The objective was to create a model to analyse how features' relevance varied among models of different experiment groups, to find groups that shared patterns. For that, the Agglomerative Clustering model was chosen to rank features using the feature importance calculated using the library ELI5, the same global explainability method used to rank features in the previous stage. Agglomerative Clustering is an unsupervised technique to group the data into the selected k number of clusters. It uses a distance function to calculate the distance among observations in the feature space. It first assigns a cluster to each observation and then it recursively merges the clusters that are near given their observations, reducing the number of clusters while augmenting their size. This procedure is continued until there only remain the selected k number of clusters. One advantage of this algorithm is that it allows to visualise the distance among observations and clusters in a hierarchical scheme named dendrogram. In addition, it enables to select the number of clusters automatically: by selecting the distance limit in height to stop merging the clusters when visualising the dendrogram.

Fig. 1. Bushing testbed.

Model explanations were analysed together with expert technicians, taking into account the theoretical and experimentally inferred mechanical knowledge from the use-case in order to test whether they were comparable or not, validating that knowledge using real data.

IV. Application Data

A. Bushings testbed

The data collected for the research of this work is based on a set of experiments performed in a fatigue testbed. Their aim is to find the characteristics under which bushings are able to tolerate the biggest accumulated load, which equals to seeking characteristics that make experiments last longer. The discovered optimal operational characteristics will be extrapolated to real machines in order to improve their components' working life. This is the reason why testbed's EOCs are similar or proportional to real ones.

The testbed consists of a hydrodynamic journal bearing and a shaft that is rotating inside it. This system is used to reduce friction between moving and static parts; with applications in switching between rotary and linear motion in big load operation environments. The gap between them is lubricated with oil, enabling the hydrodynamic state. Fig. 1 contains a picture of the testbed. The PhD dissertation by Hassasin [28] presents the underlying theory of this type of bushings and uses a testbed to apply High Frequency Vibration Analysis techniques for PdM.

Expert technicians can perform general approximate reasoning and predict when the experiment will end just few seconds before failure occurs, by monitoring sensor variables and analysing them based on domain knowledge. However, they do not understand the influence of each variable in their remaining time and neither they can predict the remaining life until last observations.

B. Dataset description

The dataset consists of 576 experiments and 97 EOC variables. Some variables are time-series data collected with a sampling rate of 1 sample per second. The remaining are process variables composed of experiment characteristics and identifiers, so they are constant for each experiment.

The collected time-series variables are from sensors related to speed, load, temperature and lubrication. Furthermore, material quality has been measured using testers.

Given the industrial nature of data, the same characteristics tested in different experiments have different results. This happens due to differences in components' manufacturing tolerances, environmental conditions, assembly adjustments etc.

The experimenting procedure consists of first choosing which characteristics to test in a set of experiments. After that, several experiments are performed using the chosen characteristics and average duration is calculated to abstract from components' manufacturing variability.

The dataset lacks of incorrect or unfinished experiments because expert technicians have removed them.

V. Results and discussion

A. Results

The visual and analytical analysis of the dataset helped understand and identify data characteristics and how fatigue experiments were performed. Moreover, a summary data dictionary was created in a table of 97 rows and 6 columns. Its columns contained the following information of the recorded variables: name, type, unit, type of feature, meaning and comments, and each row contained the information of each recorded variable.

There is the need to measure the performance of the developed models for any ML task, in order to measure the precision with which the models perform the desired task, allow comparisons among different models and assist in model optimisation in architecture selection and parameter setting.

Two common techniques to evaluate regression models are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). MAE is the normalised sum of all the absolute errors between the real and predicted values (1). RMSE is the square root of the normalised sum of all the squared errors between real and predicted values (2).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|}{n} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y_i})^2}{n}} \tag{2}$$

In the aforementioned equations, i indicates the number of observation, y_i is the real value of the variable in the observation i and $\hat{y_i}$ is the predicted value for the variable in the observation i.

The MAE metric is more intuitive to understand given that it indicates the absolute error performed by the model on observations on average. In contrast, RMSE is more sensible to outliers than MAE, being interesting for some use-cases. Therefore, RMSE has been chosen as a metric to train models so that they are more robust for outliers. For model ranking, both metrics are analysed and for model overall performance evaluation MAE is chosen, given it is easier to interpret and understand even for domain technicians.

TABLE I
Model comparison before feature selection process.

Algorithm	**RMSE**		**MAE**	
	mean	**std**	**mean**	**std**
Gaussian Naive Bayes	450.39	16.82	331.59	16.56
Linear Support Vector Regressor	421.14	208.67	300.69	137.16
K-Neighbors Regressor	258.57	20.65	163.44	13.20
Linear Discriminant Analysis	106.54	18.12	69.94	5.67
XGBoost Regressor	**53.96**	12.33	**36.10**	5.11
Random Forest Regressor	61.64	12.97	36.29	5.43

Table I shows the aforementioned machine learning regression algorithms performance of 10-fold cross-validation using the MAE and RMSE metrics.

The models that obtained best results were the tree ensemble ones, aka Random Forest and XGBoost. On average, XGBoost obtains better RMSE and slightly better MAE than Random Forest. However, given their standard deviation, we used t test to analyse whether there is enough statistical evidence for models' performance being different or not along 10 folds given MAE and RMSE metrics. The results show there is not statistical evidence on performance metrics and therefore, the criteria for choosing one model above the other was chosen to be explainability facility, another objective for the created model. Thus, Random Forest Regressor was picked over XGBoost Regressor given it is simpler and provides naive feature explanation by feature importance, which makes it is easier to interpret.

Then, the chosen Random Forest model's robustness was analysed by executing it with 20 random initialisation seeds over the mentioned 10 folds cross validation. The results show it is stable along different folds with a average mean of 35.89 and standard deviation of 5.70, given experiments duration is several hundreds of seconds and domain technicians can only estimate remaining life in experiments latter observations. Based on this results, we split the dataset in a random subset of 80% for training and 20% for testing to iterate over feature selection and rank models much faster.

After performing feature selection using expert knowledge and evaluation of feature importance for the model, the result was a *Random Forest Regressor* model executed on a subset of 10 original features that obtains a MAE of 41.29 on new split test data. Concretely, five setting features, four sensor features and an operational feature were selected. Fig. 2 shows the performance of the model during an average experiment. Table II shows the same experiments as Table I after performing feature selection, using the same 10-fold cross validation and metrics for evaluation. Given the results are similar for last

TABLE II
MODEL COMPARISON AFTER FEATURE SELECTION, SELECTING THE 10 MOST RELEVANT FEATURES.

Algorithm	RMSE		MAE	
	mean	std	mean	std
Gaussian Naive Bayes	299.25	20.50	215.83	13.28
Linear Support Vector Regressor	954.88	1332.73	770.48	1066.45
K-Neighbors Regressor	264.12	21.98	193.80	16.58
Linear Discriminant Analysis	187.56	15.86	140.57	12.71
XGBoost Regressor	**52.54**	9.68	**36.29**	4.32
Random Forest Regressor	58.36	11.86	38.25	4.80

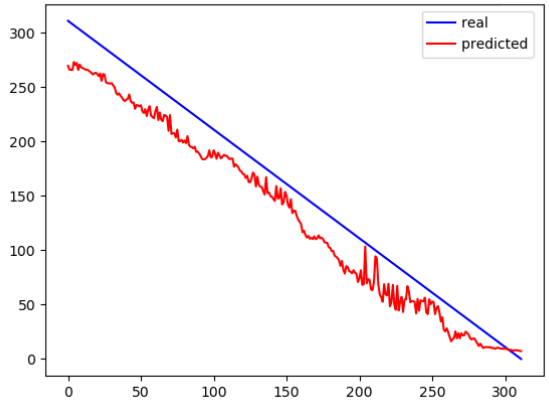

Fig. 2. Stage three's final model remaining life prediction on an average fatigue test. x axis indicates the number of observation whereas y axis indicates the remaining time in seconds. The blue line indicates the real remaining life and the red indicates the predicted one by the model.

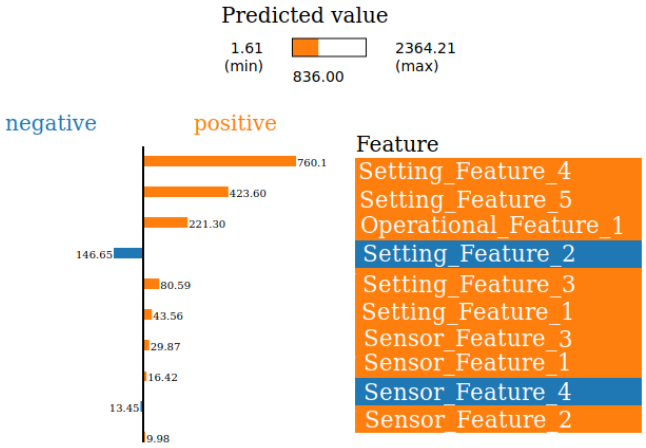

Fig. 3. Local explanation of a remaining life prediction in an experiment observation. The prediction equals the real value: 836 seconds.

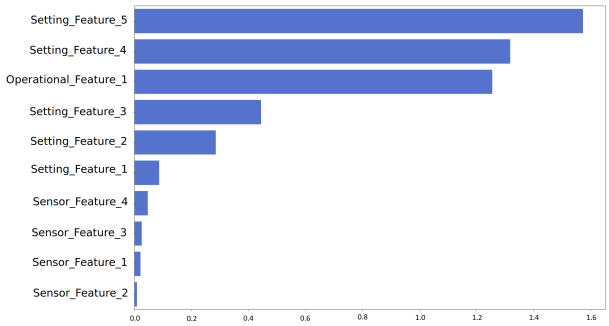

Fig. 4. Global explanation of remaining life predictor model. It shows feature importance ordered from highest to lowest. The relevance magnitude is indicated in x axis.

two models, the selected 10 feature subset keeps representative information from the original data, simplifying the problem and thus facilitating explainability.

Global explanations have been used to guide feature selection process and also to understand overall model performance by analysing feature importance. This enables feature ranking and analysis of contribution to model prediction. Moreover, this enabled to contrast model's overall performance with domain technicians conclusions given their expertise and analysis of experiment's results, which turned out to be similar. Likewise, local explanations are used to analyse model behaviour by explaining predictions of several experiments observations. Domain technicians have found local explanations application on analysing how target variable is influenced by changes in selected features. This could be used for experiments setting optimisation to achieve better results. Conversely, local explainability results could be improved given linear models may not be able to model non-linear data correctly.

Local explanations have been extracted from the previous model using the library LIME. Fig. 3 shows how each feature affects model's estimation given an experiment observation. On the top part, model's prediction for the recorded data instance is shown; in this case it is estimated to be 836. The Feature list shows names of predictor features in current observation ordered by importance. The bottom left graphic shows how each feature of previous table affects the aforementioned estimated value. The features are ordered by local importance top to bottom, where value indicates how much influence each has on current prediction and the color indicates in which direction its change affects model's prediction.

Fig. 4 shows the feature importance calculated on the aforementioned model based on *ELI5*. Global explanations were obtained using feature permutation technique, which consists of randomly shifting only the values of one feature and not modifying the rest. Then, model score changes between original and this data are analysed for each predictor shifting, being an objective way to analyse which features are more important for the model.

According to the figure, the first 5 variables concentrate most of the information the model needs, while remaining 5 have less importance.

To the already mentioned model of 10 features, automatic data-driven feature selection techniques were applied with the objective to select only the most relevant variables that still achieve an acceptable performance on the model. The feature selection techniques and their results are summarised in Table

TABLE III
FEATURE SELECTION TECHNIQUES.

Algorithm	**5 features MAE**	**10 to 3 features mean MAE**
Recursive Feature Elimination, model default feature importance, step=1	**45.74**	**57.37**
Select most relevant incrementally, ELI5 feature importance, step=1	**45.74**	**57.58**
mRMR, using MIQ, reduce step=1	109.67	78.67
kbest, mutual info	112.02	103.16
kbest, ftest	191.46	127.36
select lassoCV	176.63	146.07
select ridgereg	218.13	176.38

Feature selection techniques tested to reduce dimensionality of model from 10 to 3 features using sklearn, mRMR and XAI-based methods

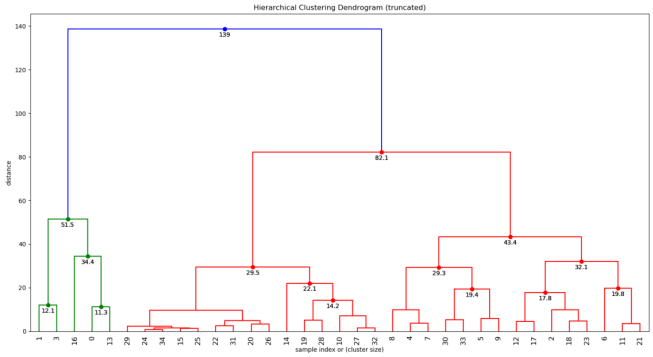

Fig. 5. Dendrogram of clustering feature importance score of models trained with data grouped by process variables.

III. The results are ordered by lowest to highest MAE of the chosen model with selected 5 features.

Finally, these are the experiments performed in the fourth stage of the work, using the data grouped under same characteristics of setting variables. First, the calculated model importance scores were used as new variables to identify the groups. Using this data, an agglomerative clustering was performed. Its results were analysed by cutting the dendrogram of Fig. 5 into different cluster sizes, from 2 to 10.

The dendrogram shows that two or three clusters could be clearly distinguished, by setting the maximum cluster distance to 100 and 60 respectively. To understand the clusters in a visual way, we created a plot using two representative variables for experiments' results. Fig. 6 shows the results of choosing a cluster size of 3.

The image shows that, in the chosen plot, the clusters are disperse. We expected to assign groups that were near to the same cluster and the ones that were far to other clusters, but this does not happen.

B. Discussion

This section discusses the results obtained after performing experiments of stages three and four of this work.

Table III discusses the outcome of comparing feature selection algorithms:

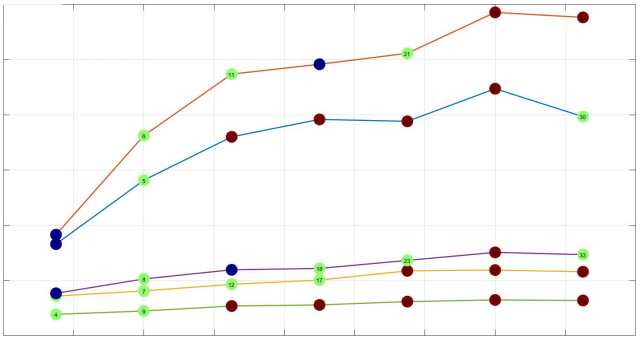

Fig. 6. Results of clustering experiments. x axis shows a feature related to fatigue and y axis shows a experiment setting variable. Each ball represents a group of experiment data and its color indicates the cluster assigned by the algorithm. The coloured lines that join these balls identify experiment characteristics.

1) The algorithms that obtain the best results are based on regression algorithm's importance metrics.
2) After them, mRMR obtains the best results.
3) Linear models are not an accurate feature selection algorithm for this problem because they are unable to model data's non-linearity.

The selected subset of variables is suitable given that the developed model is accurate. Its performance has also been visualised using plots of each experiment in test data, enabling models' performance visual analysis. After analysing the top selected features' performance, the most relevant features are setting variables. The reason is that four out of five most relevant belong to this type, and the remaining five have much less importance.

Regarding the results of applying XAI techniques to models, this paragraph discusses their suitability. Global explanations are fair because the model performs a random shift of feature values. However, local explanations might be unfair given that there are categorical variables, whose change can have high impact on model's output. Even so, global and local explainability work accordingly, weighting feature importance in a similar way.

On one hand, the third stage's model works well. Its aim is to optimise the process of fatigue tests reducing the number of experiments and inferring knowledge about which are the variables that influence remaining life the most. The next step will be its industrialisation by deploying it into the Programmable Logic Controller (PLC) of bushing testbed so technicians are able to see the predicted remaining life in real time for new fatigue tests. After that, the evolution of model's performance will be tracked and a retraining strategy will be defined.

On the other hand, the models generated for each group of experiments in fourth stage have not improved their performance over the previous general model. A possible reason could be that there is much less data available for training each of them, which makes the model difficult to generalise. Moreover, these models cannot be grouped by feature importance because, as the clustering model has shown, the generated

groups do not show any clear relation.

VI. Conclusions

All in all, the result of this work is a ML model that predicts accurately remaining life of experiments and gives global and local explanations in order to understand its predictions, being industrial process knowledge crucial for both accuracy and understandability goals.

Therefore, data scientist mindset and expert knowledge combination is the path to integrate complex data-driven models into industrial companies, resulting in hybrid models. This can be achieved by combining models with XAI techniques, taking advantage of expert knowledge to guide: model creation, interpretation and optimisation, in order to link its predictions with physical meaning. This will enable global and local understanding of model predictions, even in the case of black-box ML models.

Moreover, future research is promising to integrate explainable machine learning models to optimise, automatise and assist knowledge discovery in industrial processes based on data. This would bring the reduction in maintenance costs and improve machines availability, performance and quality.

With the mentioned objectives in mind, future research in this field could move towards obtaining an interpretability-accuracy trade-off. Accordingly, models should be accurate enough to perform the selected task and at the same time interpretable. Thus, expert technicians could infer knowledge, perform diagnosis and finally trust their predictions.

References

[1] D. Lukač, "The fourth ict-based industrial revolution" industry 4.0"—hmi and the case of cae/cad innovation with eplan p8," in *2015 23rd Telecommunications Forum Telfor (TELFOR)*. IEEE, 2015, pp. 835–838.

[2] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20 590–20 616, 2017.

[3] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Transactions on systems, Man, and Cybernetics*, no. 1, pp. 28–44, 1973.

[4] L. Liao and F. Köttig, "A hybrid framework combining data-driven and model-based methods for system remaining useful life prediction," *Applied Soft Computing*, vol. 44, pp. 191–199, 2016.

[5] Y. Li, T. R. Kurfess, and S. Y. Liang, "Stochastic prognostics for rolling element bearings," *Mechanical Systems and Signal Processing*, vol. 14, no. 5, pp. 747–762, 2000.

[6] C. H. Oppenheimer and K. A. Loparo, "Physically based diagnosis and prognosis of cracked rotor shafts," in *Component and Systems Diagnostics, Prognostics, and Health Management II*, vol. 4733. International Society for Optics and Photonics, 2002, pp. 122–133.

[7] A. Diez-Olivan, J. A. Pagan, R. Sanz, and B. Sierra, "Data-driven prognostics using a combination of constrained k-means clustering, fuzzy modeling and lof-based score," *Neurocomputing*, vol. 241, pp. 97–107, 2017.

[8] N. Kolokas, T. Vafeiadis, D. Ioannidis, and D. Tzovaras, "Forecasting faults of industrial equipment using machine learning classifiers," in *2018 Innovations in Intelligent Systems and Applications (INISTA)*. IEEE, 2018, pp. 1–6.

[9] X. S. Si, W. Wang, C. H. Hu, M. Y. Chen, and D. H. Zhou, "A Wiener-process-based degradation model with a recursive filter algorithm for remaining useful life estimation," *Mechanical Systems and Signal Processing*, vol. 35, no. 1-2, pp. 219–237, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.ymssp.2012.08.016

[10] J. L. Liang Guo, Naipeng Li, Feng Jia, Yaguo Lei, "A recurrent neural network based health indicator for remaining useful life prediction of bearings," *Boletin Tecnico/Technical Bulletin*, vol. 55, no. 16, pp. 585–590, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neucom.2017.02.045

[11] K. Pichler, E. Lughofer, M. Pichler, T. Buchegger, E. P. Klement, and M. Huschenbett, "Fault detection in reciprocating compressor valves under varying load conditions," *Mechanical Systems and Signal Processing*, vol. 70, pp. 104–119, 2016.

[12] F. Zhao, Z. Tian, and Y. Zeng, "Uncertainty quantification in gear remaining useful life prediction through an integrated prognostics method," *IEEE Transactions on Reliability*, vol. 62, no. 1, pp. 146–159, 2013.

[13] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[14] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, 2019.

[15] J. M. Alonso, C. Castiello, and C. Mencar, "A bibliometric analysis of the explainable artificial intelligence research field," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, J. Medina, M. Ojeda-Aciego, J. L. Verdegay, D. A. Pelta, I. P. Cabrera, B. Bouchon-Meunier, and R. R. Yager, Eds. Cham: Springer International Publishing, 2018, pp. 3–15.

[16] J. M. Alonso, L. Magdalena, and S. Guillaume, "Hilk: A new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism," *International Journal of Intelligent Systems*, vol. 23, no. 7, pp. 761–794, 2008.

[17] W. Samek, *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer Nature, 2019.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[19] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[21] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, 2015.

[22] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[23] J.-R. Rehse, N. Mehdiyev, and P. Fettke, "Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory," *KI-Künstliche Intelligenz*, vol. 33, no. 2, pp. 181–187, 2019.

[24] M. Carletti, C. Masiero, A. Beghi, and G. A. Susto, "Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 21–26.

[25] C. Shearer, "The crisp-dm model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[26] TeamHG-Memex, "Explain like i'm five (eli5)," https://github.com/TeamHG-Memex/eli5, 2019.

[27] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 8, pp. 1226–1238, 2005.

[28] O. A. Hassin, "Condition monitoring of journal bearings for predictive maintenance management based on high frequency vibration analysis," Ph.D. dissertation, University of Huddersfield, 2017.