IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 July 2025, accepted 17 August 2025, date of publication 21 August 2025, date of current version 4 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3601385

RESEARCH ARTICLE

Cost-Efficiency and Cost-Effectiveness of XAI in Predictive Maintenance

PETER HUGHES[1], **SURESH PERINPANAYAGAM**[1], **(Member, IEEE), AND PETER BALL**[2]
[1]School of Physics, Engineering and Technology, University of York, YO10 6DD York, U.K.
[2]School for Business and Society, University of York, YO10 5ZF York, U.K.

Corresponding author: Peter Hughes (pmh536@york.ac.uk)

ABSTRACT Predictive maintenance aims to reduce operational costs by anticipating and preventing system failures or inefficiencies. While high-performance AI models such as neural networks offer accurate predictions, their lack of transparency limits their usefulness for guiding interventions. Conversely, explainable AI (XAI) models provide insight but often at the expense of accuracy. This paper proposes a framework for cost-based evaluation of interpretable AI models in predictive maintenance, using both classification and regression contexts. We establish criteria to determine when the benefit of interpretability outweighs any reduction in accuracy and show that the utility of XAI is bounded by the relative cost of maintenance versus failure. These findings offer practical tools for assessing the business case for interpretable models in predictive maintenance and related domains. In research, the criteria enable cost-based evaluation and comparison of alternative machine learning methods for regression and classification.

INDEX TERMS Cost-effectiveness, cost-efficiency, explainable AI, predictive maintenance.

I. INTRODUCTION

The ‘Fourth Industrial Revolution’ is characterized by rapid technological advancements, paradigm-shifting changes, and transformative impacts across industries [1]. Central to this transformation is the evolution of engineering maintenance strategies that leverage technologies such as artificial intelligence (AI) and machine learning (ML) to optimise the performance of critical systems [2], [3].

AI enables organisations to predict malfunctions and optimise maintenance schedules, reducing both downtime and associated costs. However, while advanced models such as neural networks offer high predictive accuracy, they often lack transparency, making it difficult for decision-makers to understand and trust their outputs [4], [5], [6]. This lack of explainability can limit the adoption of AI in critical decision-making scenarios [7], where trust and insight are essential.

In response, interest has grown in Explainable AI (XAI), which prioritises interpretability alongside performance. Interpretable models, however, often underperform compared to black-box alternatives, leading to the perception of a ‘trade-off’ between predictive accuracy and explainability [8].

We examine this perceived trade-off between predictive performance and interpretability, proposing that these two qualities can be jointly evaluated in terms of their contribution to overall cost performance. In this view, explainability is not merely an aesthetic or regulatory concern, but a factor with measurable economic impact. We reframe the trade-off as a cost-benefit problem: interpretability must justify its value by offsetting any performance loss through improved decision-making, reduced error costs, or gains in operational efficiency.

This perspective aligns with the view that predictive maintenance (PdM) aims to reduce the “financial and time costs of upkeep” [9] through more effective and efficient maintenance. Accordingly, we evaluate AI-based PdM systems in terms of their expected financial return. In this context, we define cost-efficiency as the ability to deliver a net reduction in operating costs, and cost-effectiveness as a basis for selecting the most appropriate form of PdM among alternatives.

Two research questions are addressed to establish criteria that determine when the value of interpretability offsets a loss of predictive accuracy:

1. Using measures of predictive accuracy, what level of explanatory benefit is required for an XAI model to be cost-efficient? (RQ1)

The associate editor coordinating the review of this manuscript and approving it for publication was Yiqi Liu.

2. How can measures of predictive accuracy be used to express the relative cost-effectiveness of different models? (RQ2)

Our motivation is to develop practical, cost-based evaluation criteria to strengthen the credibility of our future investigations and provide a basis for comparison with other work. There is a growing body of work exploring the application of explainable AI in predictive maintenance across industrial sectors. In the energy domain, for example, explainable models have been used to improve the transparency of fault detection in wind turbines [10]. In aerospace, post-hoc feature significance of deep learning RUL predictors for turbofan engines has been used to relate individual sensor features estimates of degradation [11]. Similarly, in manufacturing, feature significance has been proposed to provide interpretable insights, allowing managers to align predictive recommendations with operational priorities [12]. These examples illustrate that explainability is not only a theoretical concern but has practical implications.

While we derive criteria for cost-effectiveness within the context of PdM, the underlying principles are expected to be applicable to other domains. In areas such as healthcare, finance, and risk management, the definition of cost differs, but information that supports better outcomes or reduces the consequences of errors can hold greater value than marginal gains in predictive accuracy.

Section II addresses the cost-efficiency of classification models (both binary and multi-class), and Section III considers cost-effectiveness for classifiers. Regression models play an important role in PdM, particularly for estimations of Remaining Useful Life (RUL). An approach for applying our criteria to regressors is presented in Section IV. A discussion and our conclusions follow in Sections V and VI, respectively.

II. CLASSIFICATION COST-EFFICIENCY

It is widely asserted that understanding an AI system's reasoning enhances its adoption and effective use. However, academic research rarely translates this premise into actionable criteria for real-world deployment. Implementing AI in practice involves significant investments of time, money, and resources. This section presents an interpretation of cost-efficiency for both binary and multi-class classifiers, with the goal of developing criteria to evaluate the utility of explanatory models.

The following symbols are used throughout our analysis:

- TP: The number of failures correctly identified.
- FN: Failures incorrectly predicted as non-failures.
- TN: Correctly identified non-failures.
- FP: Non-failures incorrectly identified as failures.
- f: the estimated cost of an unexpected failure.
- m: the estimated cost of maintenance to prevent a predicted failure.

To present the equations in a clear and concise manner, we define several derived quantities:

- **Precision** - the proportion of positive predictions that are accurate: $\mathrm{TP}/(\mathrm{FP}+\mathrm{TP})$
- **Recall** - the proportion of true failures that are correctly predicted: $\mathrm{TP}/(\mathrm{FN}+\mathrm{TP})$
- Falsehood ratio **(FR)** – the number of false positives as a proportion of actual failures: $\mathrm{FP}/(\mathrm{FN}+\mathrm{TP})$
- Failure Mitigation Cost Ratio (**FMCR**): the cost of maintenance to address a predicted failure prior to an explanation as a proportion of the expected cost of a failure: m/f.
- Explanatory factor, α: the ratio of expected cost of maintenance guided by an explanation to the cost of maintenance with no explanation.

A. BINARY CLASSIFICATION

For a predictive maintenance model to be cost-efficient the costs of preventative maintenance to address 'predicted' failures must be less than the costs that would have been incurred if the 'actual' failures had occurred. That is, the savings from addressing predicted failures (true positives) must outweigh the cost of unnecessary maintenance (false positives). This condition can be expressed as:

$$TP * (f - m) > FP * m$$

Rewritten as: $\mathrm{TP} * (1 - m/f) > \mathrm{FP} * m/f$

And simplified to:

$$\frac{m}{f} < \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{1}$$

In statistical terms, Equation (1) shows that a model is cost-efficient if its precision exceeds the ratio of the preventative maintenance cost to the cost of failure. This represents a conservative or worst-case scenario - one that assumes every predicted failure leads directly to maintenance action. In practice, a predicted failure is often followed by inspection before an intervention. If, λ represents the fraction of the maintenance cost required for an inspection the adjusted condition for cost-efficiency becomes:

$$\frac{m}{f} < \frac{\mathrm{TP}}{\mathrm{TP}+\lambda FP} \tag{2}$$

This adjusted threshold reflects that false positives may trigger lower-cost inspections rather than full maintenance procedures. For the purposes of further analysis in this paper, we adopt Equation (1) as a baseline indicator of cost-efficiency. This avoids a speculative assumption about organizational practices or inspection protocols and offers conservative assessments of feasibility.

The ratio *m/f* recurs in equations throughout this paper. The baseline ratio, prior to any explanatory insight, is referred to as the 'Failure Mitigation Cost Ratio' (**FMCR**).

In evaluating XAI models, we propose that explanations may reduce the estimated maintenance cost, *m,* by enabling more precise or selective maintenance. A reduction in *m* lowers the precision threshold required for cost-efficiency

(Equation 1), improving the economic viability of explanatory models.

This simple precision-based cost-efficiency condition supports two practical assessments. First, for a model with known precision, it allows us to determine the reduction in maintenance cost that explanations must enable to achieve cost-efficiency. Second, given an estimated reduction in maintenance cost, it indicates the target precision that should guide model development Both assessments provide early indicators of the viability of predictive maintenance solutions. They can assist stakeholders to determine whether investment in XAI development is likely to yield practical value.

B. MULTI-CLASS CLASSIFICATION

Equation (1) can be applied to each class within a multi-class model. In this section we use the findings from the work of Vergara et al. [13] that compared the predictive performance of various machine learning techniques to identify faults in an internal combustion engine. We focus on the cost-efficiency of the neural network (NN) and linear regression (LR) models, using confusion matrices reported by [13] (Figs. 1 and 2.)

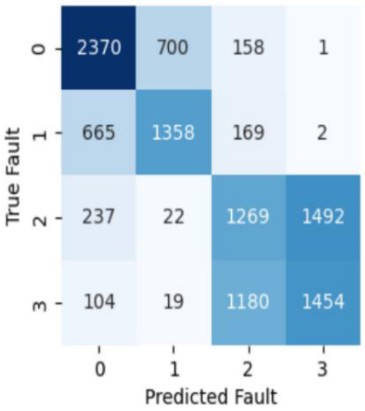

FIGURE 1. Linear regression confusion matrix from [13] for engine fault diagnosis.

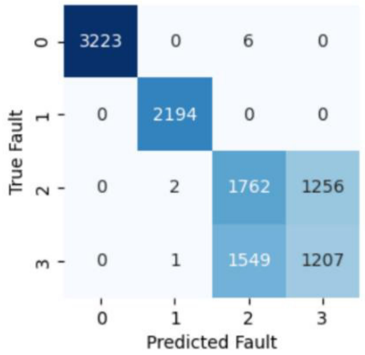

FIGURE 2. Neural network confusion matrix from [13] for engine fault diagnosis.

Examining predictions for fault 1, we see that NN clearly outperforms LR – all instances were identified and only three false predictions were made. NN's precision of 0.9986 suggests that predictive maintenance would only be cost inefficient if the cost of maintenance exceeded the cost of failure.

LR's precision of 0.6470 is significantly inferior. However, LR is a transparent algorithm and may provide actionable insights into the cause of the fault, reducing the time and resources required for repairs. If it leads to a 35.3% reduction in maintenance costs, LR would become as cost-efficient as the Neural Network despite its weaker predictive performance.

Evaluating model utility at class level provides nuanced insight into each model's strengths. A broader assessment of overall utility can be obtained by aggregating class-specific results, weighted by the prevalence of each class.

III. CLASSIFICATION COST-EFFECTIVENESS

The proposed cost-effectiveness ratio (CER) enables comparison of multiple models relative to a common benchmark. We calculate the ratio of a model's expected cost reduction to the cost reduction for a perfect black-box predictor.

For perfect prediction of a binary classification of a maintenance need, the cost reduction is the cost of failure less the cost of required maintenance for each actual failure:

$$\textit{Perfect predictor cost saving} = (\mathrm{FN} + \mathrm{TP}) * (f - m)$$

For an imperfect predictor the expected saving is:

$$\textit{Model's cost saving} = \mathrm{TP} * (f - m) - (\mathrm{FP} * m)$$

Defining our cost-efficiency ratio (CER) to be:

$$CER = \frac{\textit{Model's expected cost saving}}{\textit{Perfect predictor cost saving}} \tag{3}$$

We can state that:

$$CER = \frac{\mathrm{TP} * (f - m) - (\mathrm{FP} * m)}{(\mathrm{FN} + \mathrm{TP}) * (f - m)} \tag{4}$$

It is expected that the failure cost, f, will be consistent for both models. If we assume that maintenance costs are also consistent, division of numerator and denominator by $(f - m)$ gives:

$$CER = \frac{\mathrm{TP}}{(\mathrm{FN} + \mathrm{TP})} - \frac{\mathrm{FP}}{(\mathrm{FN} + \mathrm{TP})} * \frac{m}{(f - m)}$$

The first element $TP/(FN + TP)$ is recognised as the **recall** statistic. The quantity $FP/(FN + TP)$ represents false predictions as a proportion of the true number of classes, we have named this ratio the **falsehood ratio, FR**. Finally, the $m/(f - m)$ quantity has been rewritten in terms of the **FMCR** (m/f):

$$\mathrm{CER} = \textit{recall} - \mathrm{FR} * \frac{\mathrm{FMCR}}{(1 - \mathrm{FMCR})} \tag{5}$$

Equation 4 describes CER for a black-box model. However, if the model under test provides an explanation to reduce maintenance costs, in equation 4, the numerator's value of m will be reduced. We have added an explanatory factor (α) to represent the ratio between the reduced cost and the original cost to equation 3:

$$\text{CER} = \frac{\text{TP} * (f - \alpha m) - (\text{FP} * \alpha m)}{(\text{FN} + \text{TP}) * (f - m)} \tag{6}$$

Now, dividing numerator and denominator by $(f - m)$ gives:

$$\text{CER} = \frac{\text{TP} * (f - \alpha m)}{(\text{FN} + \text{TP}) * (f - m)} - \frac{\text{FP}}{(\text{FN} + \text{TP})} * \frac{\alpha m}{(f - m)}$$

Which, when expressing in terms of FMCR, recall and falsehood ratio, gives:

$$\text{CER} = recall * \frac{(1 - \alpha FMCR)}{(1 - \text{FMCR})} - FR * \frac{\alpha FMCR}{(1 - \text{FMCR})}$$

This can be simplified a little for:

$$\text{CER} = \frac{recall - \alpha * \text{FMCR} * (recall + \text{FR})}{1 - \text{FMCR}} \tag{7}$$

While more complex, this expression captures the trade-off between predictive performance and interpretability. The numerator is a balance of performance metrics and the explanatory factor α. However, the introduction of α as a second context-specific variable limits generalisation of individual CER values across different applications.

Equations 4 and 4.1 only apply when maintenance and failure costs differ. This is not considered a weakness as there is no business case for predictive maintenance in that scenario. A CER of greater than one is possible when explanations reduce maintenance costs sufficiently for the model to outperform perfect but unexplained predictions.

Negative CER values indicate that a model introduces more cost than benefit. The gradient of FMCR / (1 – FMCR) is positive, as FMCR increases the negative component of the denominator may dominate.

A. COST-EFFECTIVENESS ILLUSTRATION

Continuing the use of the LR and NN results from Section II-B, we evaluate the cost-effectiveness ratio for each model at a range of FMCR values, Fig 3.

The NN model is close to an optimal predictor (recall at 1.0 and only 3 false positives) and scores consistently well across the range of FMCR values.

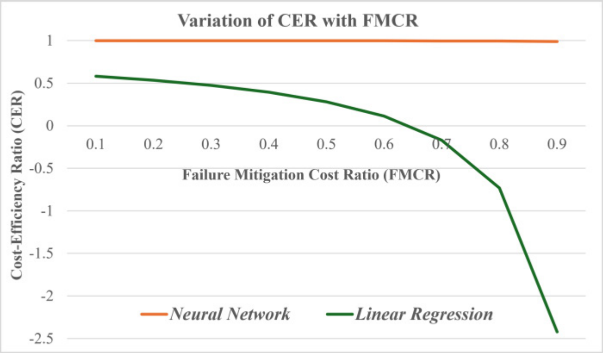

FIGURE 3. Variation of cost-effectiveness ratio with constant explanation factor ($\alpha \mathcal{D} 1$) for different failure mitigation cost ratios.

In comparison, the effectiveness of the LR model decreases significantly as FMCR increases due to the increasing cost of maintenance for false positives. At the cost-efficiency threshold of FMCR (0.647) LR's cost-effectiveness becomes negative as the cost of servicing all predicted failures now exceeds the value of preventing actual failures. We propose that **an upper limit exists for FMCR, beyond which an XAI model cannot achieve the cost-effectiveness of a more accurate predictor,** since the cost of unnecessary maintenance exceeds the benefit of correctly predicted failure prevention.

When considering how cost-effectiveness varies with explanation cost factor (α), we see a linear relationship. This is illustrated in Fig. 4 for FMCR values of 0.6, 0.5, and 0.4. Intersection of NN and LR lines indicate when LR matches NN's cost-effectiveness for values of α.

Notably, when α equals the precision of the model, the CER calculation simplifies to the recall, explaining the intersection at $\alpha = 0.647$. Additionally, as FMCR increases the influence of α becomes more pronounced, with a steeper CER-to-α gradient. The FMCR to α relationship is analysed further in Section IV-A where cost-effectiveness is illustrated for regression models.

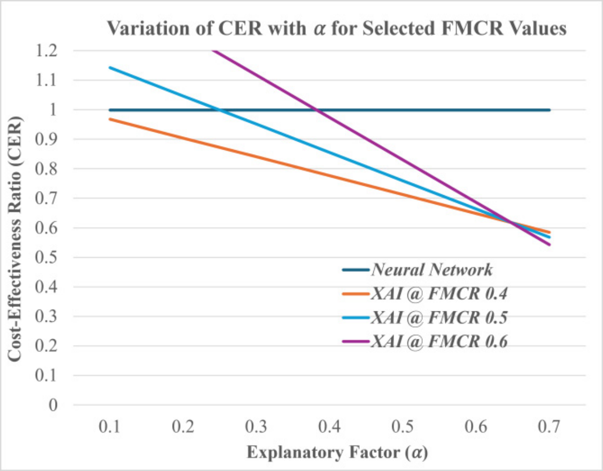

FIGURE 4. Variation of cost-effectiveness ratio with explanation factor (α) for different failure mitigation cost ratios.

IV. REGRESSION COST MEASURES

Unlike classifiers, which produce categorical outputs that can be directly mapped to misclassification costs (e.g., false positives and false negatives), regression models yield continuous predictions. In predictive maintenance, the relationship between prediction error and cost is often asymmetric and non-linear.

As a result, summary statistics used to evaluate predictive accuracy can be misleading when the aim is cost-reduction rather than predictive accuracy. For example, in predictive maintenance applications, a regression model might estimate Remaining Useful Life (RUL) for a component. Over-estimating RUL risks failure before the component is serviced, leading to unplanned downtime, safety risks, and corrective maintenance costs. Conversely, under-estimating RUL may lead to premature replacement, resulting in unnecessary maintenance costs and wasted resources. In some cases, a small over-estimate may incur greater costs than a much larger under-estimate.

While expected costs can be calculated by applying an asymmetric cost function to each test prediction and summing the results, this approach depends on a bespoke model for case-by-case estimation of explanatory benefit. Additionally, it does not address RQ2, which seeks to express relative cost-effectiveness in terms of predictive performance metrics, enabling systematic comparison of models.

To enable assessment without bespoke modelling and to address RQ2, we propose a time-window framework to reinterpret regression outcomes within a classification-like structure. In practice, maintenance operations involve lead times: planning, scheduling, and logistics often require that actions be initiated days in advance. For instance, if a predictive system is reviewed weekly to schedule servicing and order parts for the following week, then a two-week time window (covering both the current review period and the subsequent execution window) represents a meaningful operational timescale.

While existing research has emphasized prognostic accuracy metrics such as prognostic horizon and α–λ accuracy [14], these focus primarily on prediction accuracy, rather than operational utility. In contrast, our proposed time-window-based evaluation aligns with business requirements, safety regulations, and budgeting timeframes, among others.

By defining a time-window, continuous RUL predictions can be categorised: components that fail within the window are labelled as "Fail", while those surviving beyond it are labelled "OK". From this we can construct a confusion matrix comparing predictions against actual results and apply the criteria previously derived for classifiers.

In the context of predictive maintenance, a time-window representing a period of interest is considered appropriate to categorise continuous value predictions. In other domains, different criteria may be more appropriate, potentially generating multiple class categories (rather than a binary split).

A detailed, context-specific cost model may still be required to justify significant investments. However, the proposed time-window evaluation offers a practical mechanism to identify candidate models, guide further development, and reduce the burden of early-stage evaluation.

A. ILLUSTRATION OF COST ANALYSIS FOR REGRESSION

It is rare for the source data to create a confusion matrix based on a specific time window to be published alongside a paper. To illustrate the concept described above, we generated two sets of predictions: one for a more accurate model to represent a black-box predictor, and one for a less accurate model to represent an explainable AI (XAI) model. A right-skewed distribution of 'actual' Remaining Useful Life (RUL) values for 300 samples was simulated in Python, with normally distributed random errors of different magnitude applied.

Standard regression metrics for the two prediction sets (Table 1) show that the black-box model outperforms the XAI model across all regression metrics.

TABLE 1. Comparison of performance metrics.

Metric	'Black Box' Model	'XAI' Model
RMSE	4.66	10.77
MAE	3.74	8.68
R^2	0.97	0.84

To evaluate cost-effectiveness, a 14-day time window was applied to both sets of predictions, classifying components predicted to fail within the window as positive cases. Confusion matrices for each model, based on actual failures during the same period, are shown in Fig 5.

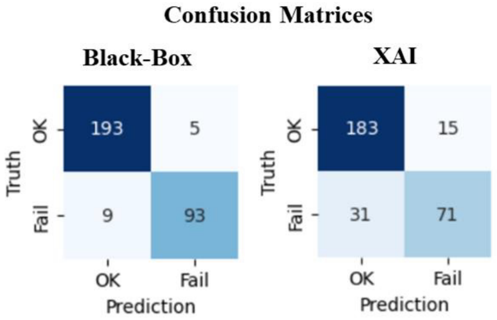

FIGURE 5. Confusion matrices for both models generated for a 14-day time window.

In calculating CER there are two context-specific variables: α (the reduction in expected reduction in cost attributable to an explanation) and FMCR (the relative cost of maintenance compared to the cost of a failure). Both these variables have a significant effect on the cost-effective ratio (CER). To investigate the FMCR-α relationship we determined the α value required for the XAI model to achieve the same cost-effectiveness as the black-box model at a range of values for FMCR. The results are plotted on Fig. 6.

As FMCR reduces, the negative term in the CER calculation reduces and a model's recall statistic dominates. Lower α values are needed for cost-effectiveness equality. When α is zero, the CER equation simplifies to $recall/(1 - FMCR)$. This represents the maximum possible cost-effectiveness an explainable model can achieve; it assumes that explanations bring maintenance costs to zero. If this value is still lower than a competing model's CER, then no level of explanation is sufficient to match a stronger predictor.

For the analysed data, if FMCR falls below 0.22, the XAI model is unable to match the CER of the black-box model. This is proposed as a general principle. The threshold value is context specific, but **for an XAI model to be more cost**

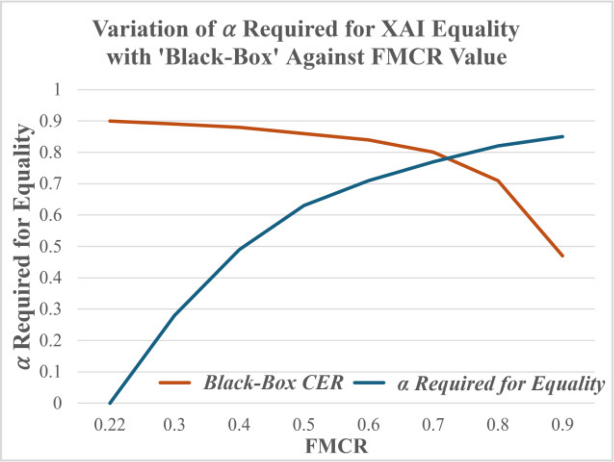

FIGURE 6. FMCR's effect on magnitude of explanatory factor α required for CER equality with a more accurate predictor.

effective than a more accurate but less interpretable alternative, there is a minimum FMCR value.

Our analysis found that for FMCR of 0.5, a 37% reduction in maintenance costs (i.e. α =0.63) is required for the explanatory model to match the black-box model. Further, with a 54% reduction in maintenance costs, the XAI model is more cost effective than a perfect black-box model; CER grew greater than 1.0.

V. DISCUSSION

It is intuitive that cost-efficiency is related to precision (how many predicted positives are true) and cost-effectiveness is related to recall (how many actual positives are identified). The criteria derived here enhance that intuition by incorporating costs and context-specific factors to define expressions that enable quantitative comparison.

For cost-efficiency, model precision is identified as an upper limit for the ratio m/f. For an XAI model, m/f can be expressed as (α ∗FMCR). This establishes a relationship for a maximum FMCR (FMCR$_{\text{max}}$) at which an XAI model can be cost-efficient:

$$\text{FMCRmax} = \frac{precision}{explanatory\ factor,\ \alpha} \tag{8}$$

Assuming that maintenance is cheaper than failure, FMCR is less than one. So, if the explanatory factor is less than precision, there is no effective restriction on FMCR; the proportion of true positives and the reduced maintenance cost ensure cost-efficiency. When α is greater than precision, equation 5 provides a clear criterion for validation of cost-efficiency using context-specific values.

This requirement for cost-efficiency, based on precision and FMCR, is complemented by the cost-effectiveness ceiling based on recall and FMCR of Section IV-A. Combined, these relationships establish upper and lower bounds for FMCR within which an XAI model is capable of comparative cost-effectiveness. Maintenance costs must be sufficiently high for actionable insights to generate significant cost savings. Equally, costs must be low enough for the cost of false positives to be outweighed by savings from prevented failures.

In Section II-A we adopted equation (1), ignoring the likelihood that false positive cost is lower than maintenance for true positives, as an inspection could identify that maintenance is not required. Applying the reasoning used for equation 1.1, we derive:

$$\text{CER} = \frac{recall - \alpha * \text{FMCR} * (recall + \lambda FR)}{1 - \text{FMCR}} \tag{9}$$

where λ is the inspection cost as a fraction of the maintenance cost. This will increase the upper FMCR limit for a model to be cost-effective. If we assume that stronger predictors generate fewer false positives, it will also improve an explanatory model's cost-effectiveness relative to its comparator.

The introduction of another context-specific variable, λ, in addition to the explanatory factor and FMCR, underscores that CER values apply to specific operational settings. This dependence on situational parameters may limit the practical utility of CER if working practices and costs vary frequently.

Finally, while these relationships provide comparative performance measures, it is important to note that absolute performance metrics depend on the consistency of class distributions between the evaluation dataset and the real-world prediction environment. Variations in class balance between these datasets can alter the confusion matrix composition, leading to misleading cost estimates. Therefore, ensuring consistent class prevalence is critical for reliable estimation of actual costs and benefits.

In summary, XAI model utility in PdM depends not only on performance metrics like precision and recall but also on contextual cost ratios and operational factors. The derived relationships provide a structured, quantifiable basis for evaluating if an XAI model can generate meaningful cost savings under real-world constraints.

VI. CONCLUSION

This study demonstrates how model precision and recall correspond to cost-efficiency and cost-effectiveness in predictive maintenance settings. In quantifying these relationships, we position XAI's explanations as cost-reducing insights and incorporate the value of explanations into criteria.

The purpose of **RQ1** was to establish a test for the feasibility of an AI model contributing to a predictive maintenance system. We found that the condition for cost-efficiency of a predictive maintenance AI model is that the expected cost of maintenance, m, relative to the cost of a failure, f, is less than the model's precision (equation 1):

$$\frac{m}{f} < precision$$

RQ2 was designed to provide a basis for comparison of candidate models. Our response is based on a calculation of cost benefit compared to an ideal predictor. A calculation of the Cost-Effectiveness Ratio (CER) enables comparison of two or more feasible AI models. This is a more complex expression (equation 7):

$$CER = \frac{recall - \alpha * \text{FMCR} * (recall + \text{FR})}{1 - \text{FMCR}}$$

where recall is the standard performance metric; α represents the proportion of maintenance cost when guided by an explanation, relative to unguided maintenance; FR is the false positive rate relative to the number of actual positives; and FMCR is a context-specific variable representing the ratio of expected maintenance cost (without explanation) to the cost of failure or breakdown.

We found that an XAI model can only match the cost-effectiveness of a better predictor if the FMCR value lies within both a lower bound (so that explanations generate significant value) and an upper bound (so that unnecessary maintenance costs do not exceed the benefit of preventing actual failures).

Time window analysis of regression models in the context of PdM is considered reasonable and enables our criteria to be applied to continuous predictions. However, a time window may not be appropriate for some domains or situations. There is scope for further research to investigate alternative cost-based evaluations of regression models.

XAI will not be suitable for all scenarios, and it is not feasible to derive global values for the applicability of XAI. The significance of explanations and operational costs are specific to contexts. Further, significant investment decisions will require consideration of costs throughout the lifecycle of a predictive maintenance system.

However, we believe the cost-based criteria we have proposed offer valuable insights and practical tools for evaluating the utility of AI alternatives, benefiting both commercial decision makers and researchers.

Organisations can apply these metrics during the planning and feasibility phase of PdM initiatives, using internal cost structures and operational parameters to assess whether the adoption of an interpretable model is justified. Future work could involve integrating these criteria into structured evaluation processes or cost-benefit templates to support early-stage investment decisions in AI-enabled maintenance systems.

REFERENCES

[1] K. Schwab, *The Fourth Industrial Revolution*. Geneva, Switzerland: World Economic Forum, 2016.

[2] T. Zonta, C. A. da Costa, R. da Rosa Righi, M. J. de Lima, E. S. da Trindade, and G. P. Li, "Predictive maintenance in the Industry 4.0: A systematic literature review," *Comput. Ind. Eng.*, vol. 150, Dec. 2020, Art. no. 106889, doi: 10.1016/j.cie.2020.106889.

[3] M. Achouch, M. Dimitrova, K. Ziane, S. Sattarpanah Karganroudi, R. Dhouib, H. Ibrahim, and M. Adda, "On predictive maintenance in Industry 4.0: Overview, models, and challenges," *Appl. Sci.*, vol. 12, no. 16, p. 8081, Aug. 2022, doi: 10.3390/app12168081.

[4] D. Gunning and D. W. Aha, "DARPA's explainable artificial intelligence program," *AI Mag.*, vol. 40, no. 2, pp. 44–58, Jun. 2019, doi: 10.1609/aimag.v40i2.2850.

[5] M. Tulio Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," 2016, *arXiv:1602.04938*.

[6] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021, doi: 10.1016/j.inffus.2021.05.009.

[7] P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, "XAI systems evaluation: A review of human and computer-centred methods," *Appl. Sci.*, vol. 12, no. 19, p. 9423, Sep. 2022, doi: 10.3390/app12199423.

[8] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 2429–2437, doi: 10.1609/aaai.v33i01.33012429.

[9] L. Cummins, A. Sommers, S. B. Ramezani, S. Mittal, J. Jabour, M. Seale, and S. Rahimi, "Explainable predictive maintenance: A survey of current methods, challenges and opportunities," *IEEE Access*, vol. 12, pp. 57574–57602, 2024, doi: 10.1109/ACCESS.2024.3391130.

[10] J. Chatterjee and N. Dethlefs, "XAI4Wind: A multimodal knowledge graph database for explainable decision support in operations & maintenance of wind turbines," 2020, *arXiv:2012.10489*.

[11] G. Youness and A. Aalah, "An explainable artificial intelligence approach for remaining useful life prediction," *Aerospace*, vol. 10, no. 5, p. 474, May 2023, doi: 10.3390/aerospace10050474.

[12] U. Dereci and G. Tuzkaya, "An explainable artificial intelligence model for predictive maintenance and spare parts optimization," *Supply Chain Anal.*, vol. 8, Dec. 2024, Art. no. 100078, doi: 10.1016/j.sca.2024.100078.

[13] M. Vergara, L. Ramos, N. D. Rivera-Campoverde, and F. Rivas-Echeverría, "EngineFaultDB: A novel dataset for automotive engine fault classification and baseline results," *IEEE Access*, vol. 11, pp. 126155–126171, 2023, doi: 10.1109/ACCESS.2023.3331316.

[14] A. Saxena, J. Celaya, E. Balaban, K. Goebel, B. Saha, S. Saha, and M. Schwabacher, "Metrics for evaluating performance of prognostic techniques," in *Proc. Int. Conf. Prognostics Health Manage.*, Denver, CO, USA, Oct. 2008, pp. 1–17, doi: 10.1109/PHM.2008.4711436.

PETER HUGHES received the B.Sc. degree in physics and the M.Sc. degree in atmospheric physics from Imperial College London, London, before a varied career in bookmaking, management, human resources, software development, architecture, and delivery. After completing the M.Sc. degree in computer science and artificial intelligence at the University of York, where he commenced a research-based Ph.D.

SURESH PERINPANAYAGAM (Member, IEEE) received the M.Eng. degree in aeronautical engineering from the Imperial College of Science, Technology and Medicine, London, U.K., in 1996, and the Ph.D. degree in aeronautical engineering from the Rolls-Royce University Technology Centre, Imperial College London, London, U.K., in 2004. He is currently a Professor of engineering with the University of York, U.K. His research interests include data-centric engineering and digital twins for intelligent systems, sensory systems, advanced reasoners, artificial intelligence, predictive capabilities, and decision-making for intelligent systems, which demonstrate self-health awareness, operational resilience, failure recovery, system configurability, and end-of-life prediction. He is an Associate Editor of IEEE Transactions on Power Electronics special issue on Robust Design and Reliability in Power Electronics.

PETER BALL received the B.Eng. degree in mechanical engineering and the Ph.D. degree in manufacturing simulation from Aston University, U.K. He is currently a Professor of operations management with the School for Business and Society, University of York. He uses modeling, simulation, and digital twins to understand performance, as well as works on practices that underpin performance. His research has been funded by U.K. Research and Innovation Bodies, including EPSRC and Innovate U.K., into environmental sustainability, circularity, multi-scale modeling, digital twins, machine learning, and systems transformation.