# To Trust or Not: Towards Efficient Uncertainty Quantification for Stochastic Shapley Explanations

Joseph Cohen[1], Eunshin Byon[2], and Xun Huan[3]

[1] *Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48109, United States*
*cohenyo@umich.edu*

[2] *Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109, United States*
*ebyon@umich.edu*

[3] *Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI 48109, United States*
*xhuan@umich.edu*

## ABSTRACT

Recently, explainable AI (XAI) techniques have gained traction in the field of prognostics and health management (PHM) to enhance the credibility and trustworthiness of data-driven nonlinear models. Post-hoc model explanations have been popularized via algorithms such as SHapley Additive exPlanations (SHAP), but remain impractical for real-time prognostics applications due to the curse of dimensionality. As an alternative to deterministic approaches, stochastically sampled Shapley-based approximations have computational benefits for explaining model predictions. This paper will introduce and examine a new concept of explanation uncertainty through the lens of uncertainty quantification of stochastic Shapley attribution estimates. The proposed algorithm for estimating Shapley explanation uncertainty is efficiently applied for the 2021 PHM Data Challenge problem. The uncertainty in the derived explanation for a single prediction is also illustrated through personalized prediction recipe plots, improving post-hoc model visualization. Finally, important practical considerations for the implementation of Shapley-based XAI for industrial prognostics are provided.

## 1. INTRODUCTION

Prognostics and health management (PHM) research has made significant strides over the last decade due to monumental improvements in data collection, storage, and computational capabilities. As elaborated by Lee et al. (2018), these advancements have helped popularize the usage of data-driven and nonlinear Industrial AI techniques for cyber-physical systems (CPS), leading to tangible improvements on key performance indicators such as yield (Senoner et al., 2022), throughput (Lai et al., 2021), sustainability (Bai et al., 2020), and reliability and safety (Xu & Saleh, 2021).

In prognostics applications, machine learning (ML) has been used both for classification (e.g., predicting incipient faults or anomalies) and regression tasks (e.g., estimating the remaining useful life (RUL) to assess the current state of degradation). In particular, deep learning models have demonstrated great success in capturing complex feature representations and handling large, high-dimensional datasets. However, they are typically presented as black-box models and lack interpretability, therefore inhibiting trustworthiness and acceptance in practice (Molnar, 2022). Motivated by regulatory, scientific, and industrial needs (Ahmed et al., 2022), explainable AI (XAI) techniques have garnered attention to address these shortcomings. This paper will focus on model-agnostic XAI motivated by Shapley analysis from cooperative game theory (Shapley, 1953), which allows for enhanced model interpretations regardless of the specific ML architecture employed to generate predictions.

Originating from game theory, Shapley explanations aim to assign the "fairest" payoff to participants in cooperative games across all possible coalitions of players (Shapley, 1953). Recently, this concept has been extended towards explaining ML predictions, typically in terms of the marginal contributions of input features (Lundberg & Lee, 2017). Computing the exact Shapley contributions is generally intractable, as the calculation scales exponentially on the number of features. As a result, a variety of approximation techniques have been developed, such as the deterministic SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017), often used in tandem with the Local Interpretable Model-Agnostic Explanations (LIME) method to ex-

plain individual predictions (Molnar, 2022). A comprehensive review of Shapley estimation algorithms can be found in (H. Chen et al., 2022).

Shapley estimation algorithms vary by model specificity, approach to removing features, and estimation strategies (H. Chen et al., 2022). For example, the L-Shapley (J. Chen et al., 2018), KernelSHAP (Covert & Lee, 2021), and Interactions-based Method for Explanations (IME) (Strumbelj & Kononenko, 2010) all compute marginal Shapley values approximations but are based on semivalue, least squares, and random order value estimations, respectively. The major limitation behind existing techniques is that their computational expense is prohibitive for applications that require analysis on a near real-time basis, such as predict-and-prevent prognostics paradigms.

Complex engineering systems require timely decisions when operating under degradation and uncertainty. While ML models can help streamline decision-making, operators may find opaque models unreliable and untrustworthy when dynamic conditions bring unexpected distributional drift from behavior represented in training sets. Therefore, obtaining fast explanations, together with the uncertainty in these explanations, can illuminate the tendencies of operational models and help decision-makers gauge the trustworthiness of the generated prognoses during operation.

While variants of SHAP have been used to obtain model explanations in industrial data, these use cases have largely been relegated to obtaining post-hoc explanations offline. For example, Senoner et al. (2022) identified quality drivers to recommend process improvements in semiconductor manufacturing, and Park et al. (2022) utilized SHAP in conjunction with other XAI approaches to improve the explainability of offline fault diagnosis for nuclear plants. Thus, there remains a clear need to develop methods that can swiftly obtain reliable model explanations with associated uncertainties.

This paper will introduce a new concept of **explanation uncertainty**, in which the user is able to quantify the uncertainty of a prediction's explanation based on the variance of stochastically estimated additive Shapley effects. The key contributions of this paper are summarized as follows:

1. Development of an algorithm to efficiently estimate Shapley-based explanation uncertainty based on the IME formulation of computing marginal Shapley values;
2. Validation using the 2021 PHM Data Challenge aerospace prognostics benchmark problem to identify key features contributing to turbofan engine failure with respect to RUL predictions; and
3. Novel visualization of Shapley-based XAI in the form of "personalized prediction recipe" plots to further enhance interpretability of predictions.

In the remainder of this paper, Section 2 will introduce the proposed algorithmic approach towards approximating explanation uncertainty, Section 3 will present a prognostics case study to demonstrate the feasibility and usefulness of the method, and Section 4 will provide concluding remarks.

## 2. METHODOLOGY

We define *explanation uncertainty* to be a quantity that represents the confidence to which the user can explain or decompose a model prediction using its inputs. Generally, explanation uncertainty is a nonlinear function of the model input, predictive model, explanation method, and respective uncertainties associated with these elements. This paper, however, directs attention to a simplified explanation uncertainty concept solely derived from a chosen explanation method: stochastic Shapley value approximations via the IME algorithm (Štrumbelj & Kononenko, 2014).

Consider an input $\mathbf{x} \in \mathbb{R}^m$ where we want to provide an explanation to its model prediction $f(\mathbf{x})$—let us call this targeted $\mathbf{x}$ the *explicand*. This $\mathbf{x}$ is associated with a Shapley explanation vector $\phi(\mathbf{x}) = \begin{bmatrix}\phi_0 & \phi_1(\mathbf{x}) & \dots & \phi_m(\mathbf{x})\end{bmatrix}^T$, where $\phi_0$ represents the baseline explanation that is constant across all $\mathbf{x}$, and $\phi_i(\mathbf{x})$ for $i = 1, 2, \dots, m$ are the marginal contributions corresponding to the $i^{\text{th}}$ feature. The basic property of local accuracy from Shapley theory then asserts

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{m} \phi_i(\mathbf{x}) \tag{1}$$

where $f$ is the original predictive model under the explanation analysis.

Originating from game theory, Shapley values are defined as the weighted average of the marginal contributions of players across all possible coalitions (Lundberg & Lee, 2017):

$$\phi_i(\mathbf{x}) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \times \left[ f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - f_S(\mathbf{x}_S) \right] \tag{2}$$

in which $F \setminus \{i\}$ represents the power set of all possible feature sets that exclude feature $i$, and the quantity $f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - f_S(\mathbf{x}_S)$ represents the effect from including feature $i$ to the subset $S$. In practice, the exact computation of a Shapley value will require $2^m$ evaluations and is infeasible for high-dimensional problems such as engineering prognostics. Due to this combinatorial complexity, Štrumbelj and Kononenko (2014) proposed the IME algorithm centered around an unbiased random order Monte Carlo (MC) estimator based on drawing $N$ samples:

$$\hat{\phi}_{i,N}(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} V_j(\mathbf{x}) \tag{3}$$

where $V_j(\mathbf{x})$ are samples drawn with replacement from a weighted distribution approximating the values of $f_{S\cup\{i\}}(\mathbf{x}_{S\cup\{i\}}) - f_S(\mathbf{x}_S)$. This is accomplished by permuting and splicing feature values from different instances, and we refer to Štrumbelj and Kononenko (2014) for the full derivation. Once the MC estimates are obtained, a local prediction $\hat{f}_N(\mathbf{x})$ can be formed for the explicand:

$$f(\mathbf{x}) \approx \hat{f}_N(\mathbf{x}) = \hat{\phi}_{0,N} + \sum_{i=1}^{m} \hat{\phi}_{i,N}(\mathbf{x}). \tag{4}$$

We extract the explanation uncertainty for the above stochastic Shapley value approximation procedure by forming the mean and variance of the $N$-sample estimator $\hat{f}_N(\mathbf{x})$:

$$\mu_{\hat{f}_N(\mathbf{x})} \equiv \mathbb{E}\left[\hat{f}_N(\mathbf{x})\right] = \mathbb{E}\left[\hat{\phi}_{0,N}\right] + \sum_{i=1}^{m} \mathbb{E}\left[\hat{\phi}_{i,N}(\mathbf{x})\right], \tag{5}$$

$$\begin{aligned}\sigma^2_{\hat{f}_N(\mathbf{x})} &\equiv \mathrm{Var}\left[\hat{f}_N(\mathbf{x})\right] \\ &= \mathrm{Var}\left[\hat{\phi}_{0,N}\right] + \sum_{i=1}^{m} \mathrm{Var}\left[\hat{\phi}_{i,N}(\mathbf{x})\right] \\ &\quad + \sum_{i=0}^{m}\sum_{j>i}^{m} 2\mathrm{Cov}\left(\hat{\phi}_{i,N}(\mathbf{x}), \hat{\phi}_{j,N}(\mathbf{x})\right).\end{aligned} \tag{6}$$

From Eq. (5), we can see that the mean of the estimator $\mathbb{E}[\hat{f}_N(\mathbf{x})] = f(\mathbf{x})$ since each $\hat{\phi}_{i,N}(\mathbf{x})$ is unbiased (i.e. $\mathbb{E}[\hat{\phi}_{i,N}(\mathbf{x})] = \phi_{i,N}(\mathbf{x})$) and so Eq. (5) reduces back to Eq. (1). Hence, the estimator $\hat{f}_N(\mathbf{x})$ is unbiased. In Eq. (6), the explanation variance requires the full covariance among all the $\hat{\phi}_{i,N}(\mathbf{x})$ since they are not independent during the IME sampling process. In the limit of large $N$, the Central Limit Theorem further implies that the distribution of the estimator $\hat{f}_N(\mathbf{x})$ converges towards a normal $\mathcal{N}\left(\mu_{\hat{f}_N(\mathbf{x})}, \sigma^2_{\hat{f}_N(\mathbf{x})}\right)$.

In our implementation, we estimate these mean and covariance terms numerically through replicating the $\hat{\phi}_{i,N}(\mathbf{x})$ approximation $K$ times and then computing their empirical mean and covariance.

## 3. Case Study

The 2021 PHM Data Challenge benchmark problem consists of synthetic run-to-failure trajectories of a small fleet of turobfan engines (Chao et al., 2021a). The engines are simulated with realistic flight conditions, and the objective is to predict the RUL of the engines (Chao et al., 2021b) given time series sensor signals. The dataset contains labeled failure mode information pertaining to five rotating components: fan, high-pressure compressor (HPC), low-pressure compressor (LPC), high-pressure turbine (HPT), and low-pressure turbine (LPT). This paper builds on previous XAI work on this dataset, which used Shapley-based explanations to derive meaningful clusters for the context of predicting future faults and the RUL (Cohen et al., 2023).

### 3.1. Results

Simple statistical features measured per flight cycle (such as minimum, first quartile, median, third quartile, maximum, mean, and standard deviation) are extracted from each time series signal. In total, 129 features are extracted based on the variables detailed in Table 1. Following data mining, the dataset is randomly split across all engine units using an 80%-20% training-testing ratio, and min-max normalization is performed and applied based on the training set. The testing set contains data from 1365 flight cycles, and will be analyzed for identifying key features explaining RUL predictions. To generate the RUL predictions, we adopt a 3-layer neural network with 64 and 32 neurons in the hidden layers and with RELU activations, and train it with the ADAM optimization algorithm. The neural network is built using Flux, a deep learning library supported by the Julia programming language (Innes, 2018). For more information for the developed predictive model and its performance, we refer to past work by Cohen et al. (2023).

Table 1. Auxiliary, operating conditions, and sensor measurement signal variable descriptions from PHM 2021 Data Challenge (Chao et al., 2021b).

| Variable | Symbol | Description | Units |
|---|---|---|---|
| $A_1$ | unit | Unit number | - |
| $A_2$ | cycle | Flight cycle number | - |
| $A_3$ | $F_c$ | Flight class | - |
| $W_1$ | alt | Altitude | ft |
| $W_2$ | Mach | Mach number | - |
| $W_3$ | TRA | Throttle-Resolver angle | % |
| $W_4$ | T2 | Total temp. at fan inlet | °R |
| $X_{s_1}$ | Wf | Fuel flow | pps |
| $X_{s_2}$ | Nf | Physical fan speed | rpm |
| $X_{s_3}$ | Nc | Physical core speed | rpm |
| $X_{s_4}$ | T24 | Total temp. at LPC outlet | °R |
| $X_{s_5}$ | T30 | Total temp. at HPC outlet | °R |
| $X_{s_6}$ | T48 | Total temp. at HPT outlet | °R |
| $X_{s_7}$ | T50 | Total temp. at LPT outlet | °R |
| $X_{s_8}$ | P15 | Total pressure in bypass-duct | psia |
| $X_{s_9}$ | P2 | Total pressure at fan inlet | psia |
| $X_{s_{10}}$ | P21 | Total pressure at fan outlet | psia |
| $X_{s_{11}}$ | P24 | Total pressure at LPC outlet | psia |
| $X_{s_{12}}$ | Ps30 | Static pressure at HPC outlet | psia |
| $X_{s_{13}}$ | P40 | Total pressure at burner outlet | psia |
| $X_{s_{14}}$ | P50 | Total pressure at LPT outlet | psia |

To obtain stochastic Shapley explanations, we utilize ShapML.jl, a relatively efficient Julia implementation of the IME algorithm that has successfully been benchmarked against other state-of-the-art Shapley estimation algorithms such as FastSHAP (Redell, 2020). To obtain the explanation uncertainty, we perform $K = 30$ replicates with sample sizes of $N \in \{10, 30, 50, 100\}$, randomizing and documenting the ShapML seed parameter in each replicate to ensure reproducibility. **A key empirical finding is that despite obtaining a unique covariance matrix for each explicand, the overall explanation uncertainty is the same for all expli-**

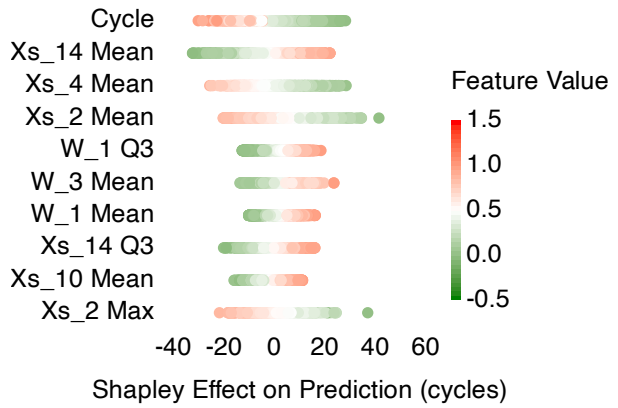


Figure 1. Beeswarm visualization ranking top 10 features in terms of mean Shapley effect, with explicands plotted and colored by normalized feature values.

**cands for a given** $N$. This has significant implications for quantifying explanation uncertainty of Shapley approximations derived from MC sampling. To the authors' knowledge, this relationship has not yet been explored in the literature and will require careful examination in future work.

Table 2 provides the runtime and explanation standard deviation for these experiments, benchmarked on a local MacBook Pro machine running macOS Ventura 13.2.1 with Apple M2 Max CPU and 32 GB of RAM. We note the efficiency of these computations, highlighting promising potential for Shapley explanations for high-dimensional prognostics applications.

Table 2. Standard deviation of a prediction's explanation and runtime for $K = 30$ replicates of $N$-sample estimators for Shapley attributions.

| **MC Samples** | $\mathbf{N = 10}$ | $\mathbf{N = 30}$ | $\mathbf{N = 50}$ | $\mathbf{N = 100}$ |
|---|---|---|---|---|
| $\sigma_{\hat{f}_N(\mathbf{x})}$ **(cycles)** | 5.39 | 3.46 | 2.61 | 1.79 |
| **Runtime (s)** | 187.50 | 624.81 | 902.27 | 1807.38 |

As performed by Senoner et al. (2022), a practical application of Shapley explanations for industrial use cases is to rank observable features in terms of mean absolute Shapley effect. In this case, this ranking provides the operator with key information and answers the following question: out of 129 collected features, which ones influence the developed data-driven RUL predictive model most? The beeswarm plot visualization in Fig. 1 serves as an informative summary ranking the top 10 most influential features, with the horizontal and color axes depicting how their distributions impact RUL predictions. The beeswarm plot is useful for succinctly illustrating global behavior and tendencies of the underlying predictive model based on the individual Shapley explanations.

A novel visualization to aid with the interpretability of individual predictions illustrating Shapley uncertainties is shown in Fig. 2. This plot depicts the "personalized prediction recipe" for a sample approaching the rated end-of-life for the engine unit. This plot is sorted by ranking mean absolute Shapley effects for a specific explicand, with the vertical axis showcasing the marginal contributions to RUL predictions relative to the estimated baseline. The error bars illustrate approximately 95% confidence intervals (two standard deviations).

Finally, the explained RUL prediction is represented by adding all the Shapley estimates and baseline, as in Eq. (4). The confidence interval of the explanation is derived from the overall explanation uncertainty shown in Eq. (6), also reported at the approximate 95% confidence level ($\pm 2\sigma_{\hat{f}_N(\mathbf{x})}$).

### 3.2. Discussion

Quantifying explanation uncertainty with the technique developed in this paper is helpful to clearly understand the impact of random variation on stochastic Shapley explanations. Because of the computational effort required to calculate Shapley effects deterministically, estimating the effects with a random order approach has distinct computational benefits, particularly when it is possible to perform uncertainty quantification.

The insights revealed in the case study are intuitive, but may also be surprising for practitioners. For example, in Fig. 1, the top 3 features include the current cycle (high feature values correlate with lower RUL), mean total pressure from the LPT outlet (high feature values correlate with higher RUL), and mean total temperature at LPC outlet (high feature values correlate with lower RUL). While some of the relationships in this plot are expected, 3 of the top 10 features included are from the scenario descriptor variables (mean and third quartile of the altitude as well as mean throttle-resolver angle).

This, in addition to the personalized prediction recipe pictured in Fig. 2, are excellent examples of XAI illuminating potentially undesirable traits of underlying models. XAI raises fundamental questions: should we trust models that are so heavily impacted by potentially noncausative features? Although we are able to generate more accurate predictions with this information, are those models useful in practice? Practitioners must be careful in interpreting Shapley explanations: all explanations are based on the dataset, model chosen, and the biases—explicit or implicit—underneath these structures. Shapley explanations merely help demystify predictions made from a black-box model, and additional experimentation is required to identify causal relationships.

The method developed in this paper aims to quantify explanation uncertainty limited to the variance of relatively efficient stochastic approximations of Shapley effects. Using this technique, the user can quantify the overall explanation uncertainty based on the fundamental property of local accuracy

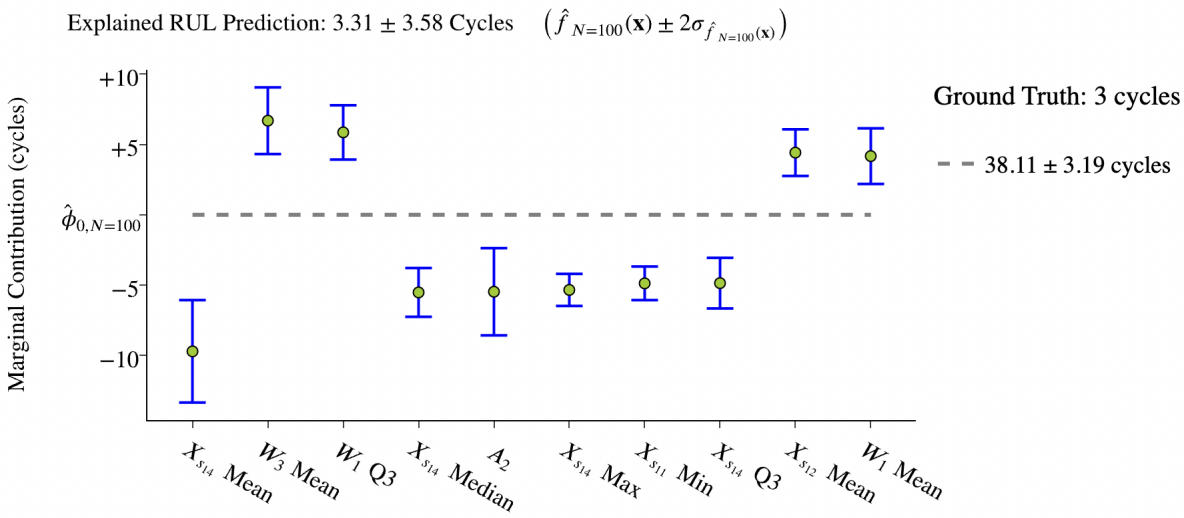


Figure 2. Personalized prediction recipe plot of a data sample approaching the engine's end-of-life. Adding the means and covariance elements of the marginal contributions (including the other 119 not pictured) plus the baseline will linearly approximate the prediction and explanation uncertainty. The mean pressure from LPT outlet, mean throttle-resolver angle, and third quartile of the altitude signals have the greatest absolute mean effect on the prediction for ths sample.

from Shapley theory. Fascinatingly, for a given number of $N$ MC samples, summing the Shapley estimates' variances and covariances will lead to an empirically constant explanation uncertainty regardless of which explicand is examined. Future work must further examine this finding, and consider other sources of uncertainties to fully tackle the problem of trustworthy data-driven modeling under uncertainty.

## 4. Conclusion

Shapley explanation theory is gaining in popularity to help explain black-box data-driven models. The simplicity of additive explanations is desirable for industry practitioners, but the computational expense of deterministic estimation methods may not be practical for high-dimensional problems. Quantifying the uncertainty of stochastic approximations of Shapley values is key to trusting explanations, and by extension, their respective models. To summarize its contributions, this paper developed:

1. An approach to efficiently estimate the explanation uncertainty purely derived from the random variation of estimating Shapley values;
2. Implementation of the explanation uncertainty concept to a high-dimensional use case, empirically finding that the total explanation uncertainty is constant across explicands for a given IME estimator; and
3. Personalized prediction recipe plots to help illustrate the uncertainties of the estimated marginal Shapley effects for a single additive explanation.

We believe these findings will encourage further contributions at the intersection of XAI and prognostics. Improved explainability will allow for more transparent, trustworthy, and effective industrial decision-making powered by big data.

## Acknowledgment

This project is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program.

## References

Ahmed, I., Jeon, G., & Piccialli, F. (2022). From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics*, *18*(8), 5031-5042. doi: 10.1109/TII.2022.3146552

Bai, C., Dallasega, P., Orzes, G., & Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. *International Journal of Production Economics*, *229*, 107776. doi: 10.1016/j.ijpe.2020.107776

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2021a). Aircraft Engine Run-To-Failure Dataset Under Real Flight Conditions. *NASA Ames Prognostics Data Repository*.

Chao, M. A., Kulkarni, C., Goebel, K., & Fink, O. (2021b). PHM Society Data Challenge 2021. *PHM Society*, 1-6.

Chen, H., Covert, I. C., Lundberg, S. M., & Lee, S.-I. (2022). *Algorithms to estimate Shapley value feature attributions.* Preprint, arxiv:2207.07605.

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). *L-Shapley and C-Shapley: Efficient Model Interpretation for Structured Data.* Preprint, arxiv:1808.02610.

Cohen, J., Huan, X., & Ni, J. (2023). *Shapley-based Explainable AI for Clustering Applications in Fault Diagnosis and Prognosis.* Preprint, arxiv:2303.14581.

Covert, I., & Lee, S.-I. (2021). Improving KernelSHAP: Practical Shapley Value Estimation Using Linear Regression. In A. Banerjee & K. Fukumizu (Eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (Vol. 130, pp. 3457–3465).

Innes, M. (2018, 5). Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, *3*. doi: 10.21105/JOSS.00602

Lai, X., Shui, H., Ding, D., & Ni, J. (2021). Data-driven dynamic bottleneck detection in complex manufacturing systems. *Journal of Manufacturing Systems*, *60*, 662-675. doi: 10.1016/j.jmsy.2021.07.016

Lee, J., Davari, H., Singh, J., & Pandhare, V. (2018). Industrial Artificial Intelligence for industry 4.0-based manufacturing systems. *Manufacturing Letters*, *18*, 20-23. doi: 10.1016/j.mfglet.2018.09.002

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).

Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book

Park, J. H., Jo, H. S., Lee, S. H., Oh, S. W., & Na, M. G. (2022). A reliable intelligent diagnostic assistant for nuclear power plants using explainable artificial intelligence of GRU-AE, LightGBM and SHAP. *Nuclear Engineering and Technology*, *54*(4), 1271-1287. doi: 10.1016/j.net.2021.10.024

Redell, N. (2020). *ShapML.jl: A Julia package for interpretable machine learning with stochastic Shapley values.* Github. Retrieved from https://github.com/nredell/ShapML.jl

Senoner, J., Netland, T., & Feuerriegel, S. (2022). Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing. *Management Science*, *68*(8), 5704-5723. doi: 10.1287/mnsc.2021.4190

Shapley, L. S. (1953). A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (p. 307-318). Princeton: Princeton University Press. doi: 10.1515/9781400881970-018

Strumbelj, E., & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications Using Game Theory. *Journal of Machine Learning Research*, *11*, 1–18.

Štrumbelj, E., & Kononenko, I. (2014). Explaining Prediction Models and Individual Predictions with Feature Contributions. *Knowledge and Information Systems*, *41*(3), 647–665. doi: 10.1007/s10115-013-0679-x

Xu, Z., & Saleh, J. H. (2021). Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliability Engineering & System Safety*, *211*, 107530. doi: 10.1016/j.ress.2021.107530