Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen

doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P1555-cd

Assessment of the Trustworthiness of Grey-Box Models for Condition Monitoring of Industrial Components and Systems

Giovanni Floreale

Energy Department, Politecnico di Milano, Italy. E-mail: giovanni.floreale@polimi.it

Piero Baraldi

Energy Department, Politecnico di Milano, Italy. E-mail: piero.baraldi@polimi.it

Francesco Di Maio

Energy Department, Politecnico di Milano, Italy. E-mail: francesco.dimaio@polimi.it

Joaquin Figueroa

Energy Department, Politecnico di Milano, Italy. E-mail: joaquineduardo.figueroa@polimi.it

Fosco Eugenio Quadri

Energy Department, Politecnico di Milano, Italy. E-mail: foscoeugenio.quadri@mail.polimi.it

Enrico Zio

MINES Paris-PSL, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France. Energy Department, Politecnico di Milano, Italy. E-mail: enrico.zio@polimi.it

In this paper we propose a method to assess the trustworthiness of Grey-Box (GB) models for Condition Monitoring (CM) of industrial components and systems. We consider a GB architecture that leverages the strengths of physics-based White Box (WB) and data-driven Black Box (BB) models by using the BB to correct the WB prediction. Specifically, the BB receives in input the measured signals and the output of the WB. We define trustworthiness in relation to model accuracy and consistency with the laws of physics. The latter is evaluated using Shapley Additive exPlanations (SHAP) to quantify the extent of reliance of the BB on the WB output. The rationale is that if the BB is sensitive to the WB output, the GB is trustworthy because it indirectly embeds the laws of physics. The effectiveness of the proposed method is demonstrated on a synthetic case study which mimics the condition monitoring of an industrial system.

Keywords: Condition Monitoring, Neural Network, Grey-box Model, eXplainable Artificial Intelligence, Shapley Additive exPlanations.

1. Introduction

Condition Monitoring (CM) relies on sensors that collect real-time data from a component or system and on models to process the data to extract relevant information on the health and operation state of the component or system under monitoring (Zio 2022). For practical use, CM of industrial components and systems requires trustworthy models, i.e. accurate and consistent with the laws of physics.

Grey-Box (GB) models, which combine a physics-based White Box (WB) model with a data-driven Black-Box (BB) model, have been proposed for CM of industrial components and systems, such as nuclear power plants (Miqueles et al. 2023), internal combustion engines (Jung 2022), dc-link capacitors (Zhao et al. 2023), data centers (Asgari et al. 2021), oil-immersed transformers in power distribution networks (Blomgren et al. 2023), chemical, petroleum and energy systems (Zendehboudi, Rezaei, and Lohi 2018).

The WB ensures transparency through its reliance on the laws of physics (Pasman and Rogers 2018);

nevertheless, simplifying assumptions are often needed to ensure timely computations, ending up in lack of accuracy. To enhance accuracy without increasing computational burden, a BB is employed to correct the WB output, thereby improving the overall accuracy of the GB (Chao et al. 2019). In this work we consider a GB architecture in which the BB, fed by signal measurements and the output of the WB, estimates the correction to be applied to the WB. Yet, despite the noticeable advantages of GB, the lack of transparency of the BB undermines their acceptability for the practical use of CM (Pensado et al. 2024).

To address this issue, the present work proposes a logic approach to assess the trustworthiness of GB, as follows: when the output of the BB is small, the GB is considered trustworthy because its estimate is mainly driven by the laws of physics at the basis of the WB; on the contrary, when the output of the BB is large, the trustworthiness of the GB estimate requires further evaluation. For further evaluation, we propose the use of SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) to assess the influence of the BB inputs (signal measurements and WB output) on the BB output (correction to be applied to the WB output). Specifically, when the BB output is not significantly influenced by the WB output, the GB is deemed untrustworthy due to its limited reliance on laws of physics; on the other hand, if the correction provided by the BB is significantly influenced by the WB output, we consider the GB output trustworthy, as it is indirectly accounting for physics.

The proposed method is demonstrated on a synthetic case study (Yang et al. 2017). The method is shown able to identify trustworthy estimates of the GB where errors tend to be smaller than those on untrustworthy estimates.

The remaining of the paper is organized as follows: in Section 2, we develop GB for CM; in Section 3, we describe the proposed method for analyzing the trustworthiness of the GB output; in Section 4, we present the synthetic case study and the accuracy of the GB; in Section 5, we discuss the obtained results in terms of trustworthiness. Finally, Section 6 draws conclusions and highlights limitations and future directions of the research.

2. GB models for CM

The objective is to develop a CM model to estimate a quantity of interest y on the basis of n measured signals $(x_1, \ldots, x_n)$. This is done using the GB model shown in Figure 1. The GB is composed of a WB and a BB. The WB receives in input the n signal measurements $(x_1, \ldots, x_n)$ and compute the estimate y_{WB} of the quantity of interest y. Then, the BB estimates the correction $\hat{c}$ to be applied to y_{WB} to match y. Finally, the output of the GB is obtained as $\hat{y}_{GB} = y_{wb} + \hat{c}$.

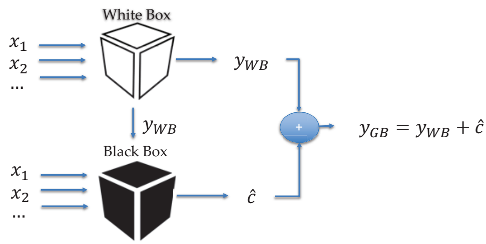

Figure 1. Considered grey-box model architecture.

The training set of the BB is made of N input $(x_1^i, \ldots, x_n^i, y_{WB}^i)$ - output (c^i) patterns, where the WB is used to compute y_{WB}^i and the error $c^i = y^i - y_{WB}^i$. The training data can be:

A) signal measurements;
B) generated by simulations from a high-resolution model.

Case A) requires that inputs and output are directly measurable and enough data are available. In some applications, e.g. involving new system designs, rare accident scenarios or harsh environments (which may prevent sensors installation for measuring input and output), A) is not feasible and B) can be applied to simulate the training data. Notice that, because high-resolution simulations are computationally intensive, they cannot serve as WBs in GBs, necessitating the use of low-fidelity alternatives (Miqueles et al. 2025).

3. Trustworthiness of the GB model

A trustworthy model is expected to be accurate and based on the laws of physics (Pasman and Rogers 2018). For the GB in Figure 1, this holds if at least one of the following two conditions on the correction $\hat{c}$, assumed to make the estimate accurate, is verified:

1) the correction $\hat{c}$ provided by the BB to the WB is small, which implies that the output of

the GB is accurate and its value mainly determined by the laws of physics of the WB;

2) the correction $\hat{c}$ provided by the BB is large, but such to make the output of the GB accurate, and driven mainly by the physics-based output provided by the WB.

Concerning the verification of 2), we employ SHAP to quantify the contribution of the input signals to the BB output. SHAP is a tool derived from cooperative game theory that fairly distributes the total gain (or "payout") from a coalition of players to each player, based on their individual contributions (Shapley 1971). Shapley values can be used to measure the influence of the input features on the output (Lundberg and Lee 2017). Specifically, in this work we are interested in the contribution of the WB output $\sigma_{y_{WB}}$ to the BB output. When $\sigma_{y_{WB}}$ is large, we can assume that the BB relies on the laws of physics.

The trustworthiness of the output of the GB can, then, be assessed considering the four zones shown in Figure 2, where the y-axis is the output $\hat{c}$ of the BB and the x-axis is the SHAP $\sigma_{y_{WB}}$ of the WB.

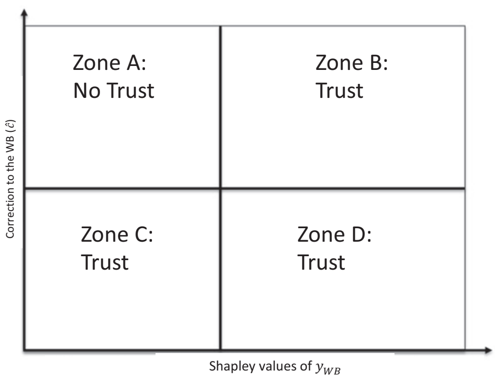

Figure 2. The four zones used to assess the trustworthiness of the GB estimations.

The only zone in which the GB output is not trustworthy is zone A, where the correction of the BB $\hat{c}$ is large but not strongly driven by the output of the WB, as indicated by the small value of the SHAP $\sigma_{y_{WB}}$. Notice that the proposed method can be easily applied in real-time due to the reduced computational effort of WB, BB and SHAP.

4. Synthetic Case Study

To demonstrate the method for the assessment of the trustworthiness GBs for CM, we consider the synthetic case study proposed in (Yang et al. 2017). The ground truth function f defining the relationship between the input signals (x_1, x_2) and the output y is:

$$y = f(x_1, x_2) =$$

$$2 + 0.01(x_2 - x_1^2)^2 + (1 - x_1) + + 2 * (2 - x_2)^2 + 7\sin\left(\frac{x_1}{2}\right) * \sin\left(\frac{7x_1x_2}{10}\right) \quad (1)$$

$$\text{with } (x_1, x_2) \in [0,5] \times [0,5].$$

Figure 3 shows y as a function of x_1 and x_2 in the considered domain.

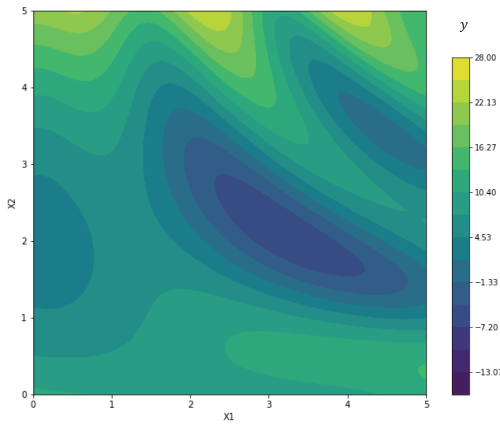

Figure 3. Function *f*.

As in (Yang et al. 2017), we assume that the function is unknown, but there is a WB which provides in output an estimate y_{WB} of y:

$$y_{WB} = 2 + 0.01(x_1 - x_1^2)^2 + (1 - x_1) + 2 * (2 - x_2)^2 + + 7\,sin\left(\frac{x_1}{2}\right) * sin\left(\frac{7x_1x_2}{10}\right) - 0.4x_1\,sin(2x_1)\,cos(x_2) \quad (2)$$

Notice that the WB model tends to significantly overestimates y for large values of x_1 and x_2 as shown in Figure 4.

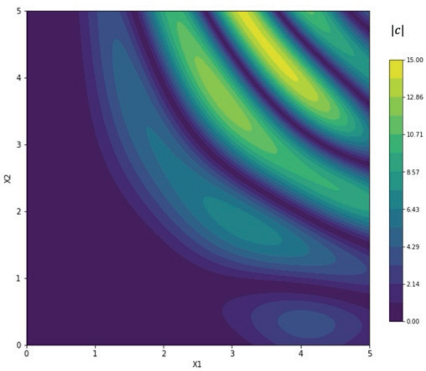

Figure 4. Absolute value of the error of the WB.

The BB is a three-layers, fully connected Neural Network (NN), with 64 neurons per layer. The training set is made of $N = 10^6$ input-output patterns obtained by randomly sampling from uniform distributions the pairs $(x_1^i, x_2^i)_{\,i=1,\dots,N}$ and the corresponding output y^i. In the test phase, the GB output is computed as $\hat{y}_{\text{GB}} = y_{WB} + \hat{c}$.

5. Results

The accuracy of the considered GB is compared with a standalone BB model, which directly estimates y from x_1 and x_2. Table 1 reports the Root Mean Square Error (RMSE) on a test set of patterns not used for training the standalone BB and the BB of the GB. The obtained results show that: *i*) the WB error is significantly larger than that of the models employing BBs and *ii*) the GB outperforms the standalone BB.

Table 1. Comparison of the accuracy of the WB, the standalone BB and the developed GB

	WB	Standalone BB	GB
RMSE	5.06	0.14	0.08

To investigate the trustworthiness of the GB, we compute the SHAP $\sigma_{y_{WB}}$ and the magnitude of the correction $|\hat{c}|$ provided by the BB for a set of $N_{test} = 1000$ test patterns. Figure 5 shows the test patterns mapped in the four zones of Figure 2. These zones are defined by setting the threshold value for $|\hat{c}|$ equal to the median of $|\hat{c}|$ in the training set and the threshold for $\sigma_{y_{WB}}$ equal to 0.33, which represents the contribution that y_{wb} would give if all three BB inputs had the same importance.

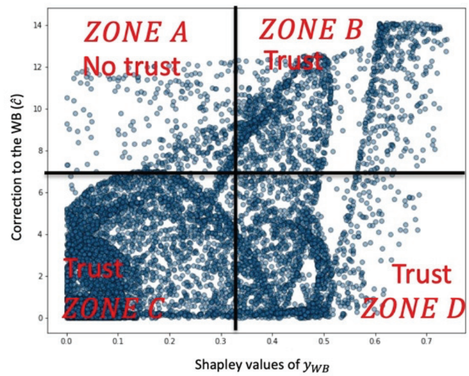

Figure 5. Projection of the test patterns in the four zones used to assess the trustworthiness in Figure 2.

The majority of the test patterns (96%) falls within the trustworthy zones B, C and D, indicating that the developed GB model is effectively consistent with the laws of physics. Figure 6 shows the test patterns in the space defined by x_1 and x_2, with the background colour representing y. The untrustworthy patterns of zone A are shown in red, the trustworthy patterns of zone B, which are characterized by large BB corrections and large $\sigma_{y_{WB}}$, are shown in orange, whereas the remaining trustworthy patterns of zones C and D are in white.

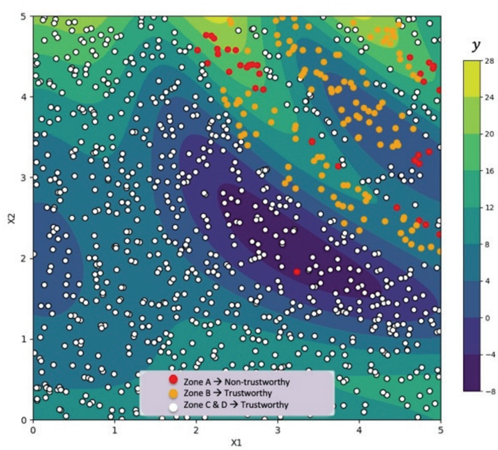

Figure 6. Scatterplot of the test patterns. The background colour indicates the true value of y, whereas the colour of the dots indicates the zones used to assess the trustworthiness (Figure 2).

Notice that the untrustworthy patterns correspond to high values of x_1 and x_2, which may be related to a difficulty of the BB model of interpolating at the boundary of the training domain. Also, the RMSE obtained on the untrustworthy patterns of Zone A (0.12), is larger than the overall RMSE of the GB (0.08). This result indicates that when the GB model does not effectively incorporate the laws of physics, it tends to produce relatively large errors.

6. Conclusions

In this work, we have developed a method to assess the trustworthiness of a GB in which the BB receives in input the output of the WB to provide in output the correction to be applied to the WB output. The proposed method is based on the use of SHAP for estimating the contribution of the inputs of the BB to its output, i.e. the correction to the WB output. Trustworthiness is assessed considering four zones with respect to the contribution to the overall GB output of the WB output and the correction provided by the BB. The results obtained on a synthetic case study show that the GB is more accurate than a WB and a standalone BB. Also, the proposed method can identify the trustworthy estimates of the GB, on which the error tends to be smaller than that on untrustworthy estimates.

A limitation of the proposed method is the need to set threshold values for defining the four zones of trustworthiness. Further investigation is needed to identify the procedure for setting them. Also, the proposed method will be applied for CM model used in real components and systems.

Acknowledgements

The participation of Giovanni Floreale and Enrico Zio to this work is supported by the SAFEPOWER project under the HORIZON-CL5-2024-D3-01 call of the European Union, Grant Agreement 101172940. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

The participation of Piero Baraldi to this work is supported by FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, Investment 1.3, Line on Artificial Intelligence).

References

Asgari, Sahar, Seyed Morteza MirhoseiniNejad, Hosein Moazamigoodarzi, Rohit Gupta, Rong Zheng, and Ishwar K. Puri. 2021. 'A Gray-Box Model for Real-Time Transient Temperature Predictions in Data Centers'. *Applied Thermal Engineering* 185. https://doi.org/10.1016/j.applthermaleng.2020.116319.

Blomgren, E. M.V., F. D'Ettorre, O. Samuelsson, M. Banaei, R. Ebrahimy, M. E. Rasmussen, N. H. Nielsen, A. R. Larsen, and H. Madsen. 2023. 'Grey-Box Modeling for Hot-Spot Temperature Prediction of Oil-Immersed Transformers in Power Distribution Networks'. *Sustainable Energy, Grids and Networks* 34. https://doi.org/10.1016/j.segan.2023.101048.

Chao, Manuel Arias, Chetan Kulkarni, Kai Goebel, and Olga Fink. 2019. 'Hybrid Deep Fault Detection and Isolation: Combining Deep Neural Networks and System Performance Models'. *International Journal of Prognostics and Health Management* 10 (4). https://doi.org/10.36001/ijphm.2019.v10i4.2621.

Jung, Daniel. 2022. 'Automated Design of Grey-Box Recurrent Neural Networks for Fault Diagnosis Using Structural Models and Causal Information'. In *Proceedings of Machine Learning Research*. Vol. 168.

Lundberg, Scott M., and Su In Lee. 2017. 'A Unified Approach to Interpreting Model Predictions'. In *Advances in Neural Information Processing Systems*. Vol. 2017-December.

Miqueles, Leonardo, Ibrahim Ahmed, Francesco Di Maio, and Enrico Zio. 2023. 'A Grey-Box Digital Twin-Based Approach for Risk Monitoring of Nuclear Power Plants'. In . https://doi.org/10.3850/978-981-18-5183-4_s24-05-579-cd.

Miqueles, Leonardo, Ibrahim Ahmed, Francesco Di Maio, and Enrico Zio. 2025. 'Importance Sampling for Monte Carlo Dynamic Event Tree Analysis of Accident Scenarios in New-Generation Nuclear Power Plants'. *Under Review*.

Pasman, Hans, and William Rogers. 2018. 'How Trustworthy Are Risk Assessment Results, and What Can Be Done about the Uncertainties They Are Plagued With?' *Journal of Loss Prevention in the Process Industries*. https://doi.org/10.1016/j.jlp.2018.06.004.

Pensado, Osvaldo, Patrick LaPlante, Michael Hartnett, and Kenneth Holladay. 2024. 'Regulatory Framework Gap Assessment for the Use of Artificial Intelligence in Nuclear Applications'.

Shapley, Lloyd S. 1971. 'Cores of Convex Games'. *International Journal of Game Theory* 1 (1). https://doi.org/10.1007/BF01753431.

Yang, Zhuo, Douglas Eddy, Sundar Krishnamurty, Ian Grosse, Peter Denno, Yan Lu, and Paul Witherell. 2017. 'Investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing'. In *Proceedings of the ASME Design Engineering Technical Conference*. Vol. 2B-2017. https://doi.org/10.1115/DETC201767794.

Zendehboudi, Sohrab, Nima Rezaei, and Ali Lohi. 2018. 'Applications of Hybrid Models in Chemical, Petroleum, and Energy Systems: A Systematic Review'. *Applied Energy*. https://doi.org/10.1016/j.apenergy.2018.06.051.

Zhao, Zhaoyang, Haitao Hu, Zhengyou He, Weiguo Lu, Herbert Ho Ching Iu, Frede Blaabjerg, and Pooya Davari. 2023. 'A Transient-Modeling-Based Grey-Box Method for Online Monitoring of DC-Link Capacitors'. *IEEE Transactions on Power Electronics* 38 (11). https://doi.org/10.1109/TPEL.2023.3308115.

Zio, Enrico. 2022. 'Prognostics and Health Management (PHM): Where Are We and Where Do We (Need to) Go in Theory and Practice'. *Reliability Engineering and System Safety* 218. https://doi.org/10.1016/j.ress.2021.108119.