Dependable Cyber-Physical Systems (DCPS) Laboratory

# Explainable Predictive Maintenance is Not Enough: Quantifying Trust in Remaining Useful Life Estimation

Ripan Kumar Kundu and Khaza Anuarul Hoque

University of Missouri-Columbia
Missouri, USA

*Annual Conference of the Prognostics and Health Management Society 2023*
*Salt Lake City, Utah, USA*

# Outline

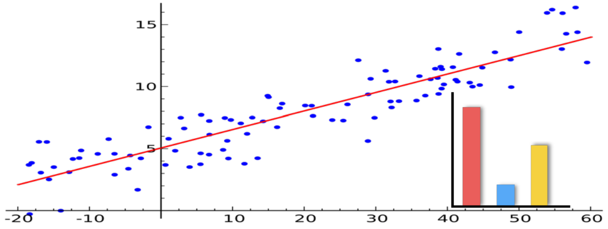
- Introduction
- Related works and their limitations
- Proposed methodology for trustworthy PdM
  - RUL local explanation methods
  - Explanation evaluation metrics
  - Robust rank aggregation and trust score measure
- Experimental results
- Conclusion & future works

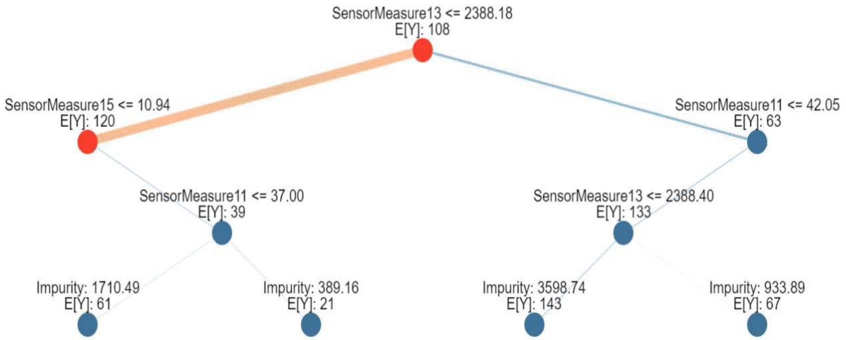
# Motivation

- Black-box machine learning (ML)/deep learning (DL) has shown tremendous success in data-driven predictive maintenance (PdM).
- It is difficult for human experts to understand and act upon black-box PdM models' decisions.
- Explanations help improve the model's understanding and provide insight into why and how the model arrived at a specific decision.
- The state-of-the-art explanation methods often suffer from the **disagreement problem**.
  - Multiple explainable AI (XAI) methods **do not agree** with a model's feature ranking.
  - Misguide the required insights by the operators and technicians to understand **what** and **why** it is happening, and **how** to react.
  - May lead to **catastrophic consequences** in safety-critical applications.
- Raise a fundamental question: **how to choose the correct explanation method for PdM models?**

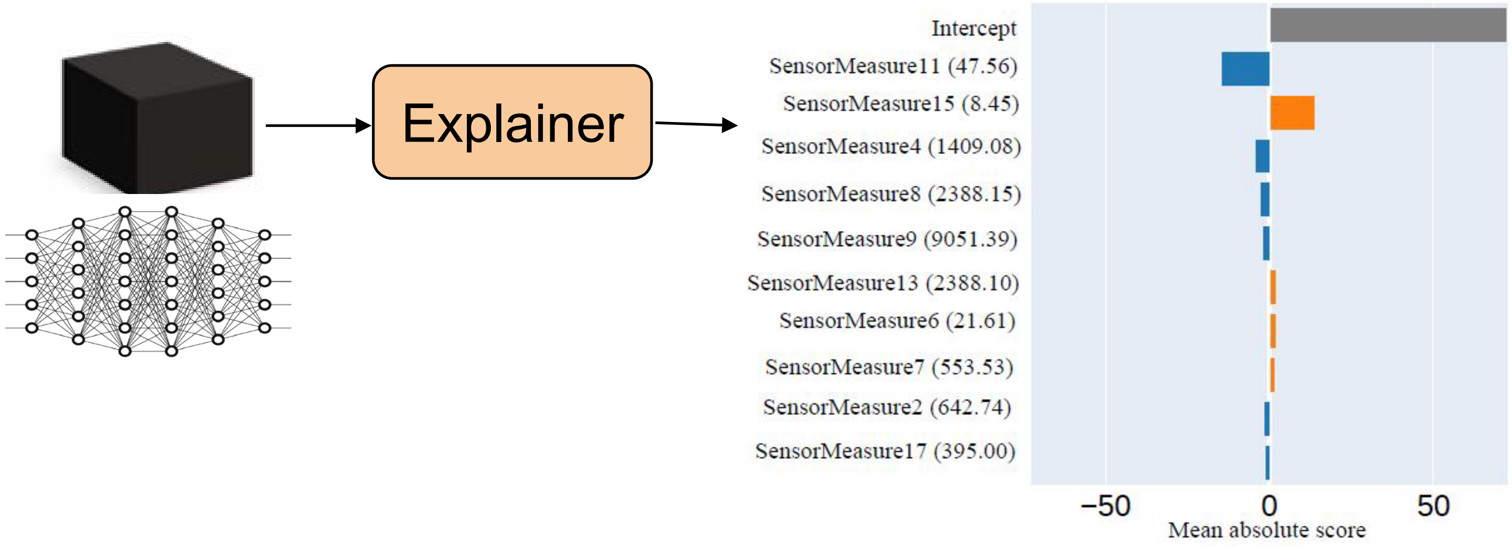
# Related works

**Take 1**: Build inherently interpretable predictive models (e.g., Rule Based Models, Generalized Additive Models, etc.,) [3]






**Take 2**: Explain pre-built models in a post-hoc manner (e.g., SHAP, LIME, etc.,) [4,5]




- Only a few works exist when it comes to evaluating the quality of the explanation of PdM models [3,5]
  - Stability and consistency
- **No work** on how to choose an **accurate** and **trustworthy** explanation for explaining the predictive RUL.
- **Unstable** and **inconsistent** explanations may lead to an **untrustworthy** PdM model for the end-users.

# XAI Limitation in PdM

- For a single prediction, the local explanations are chosen when there is a disagreement between the SHAP and LIME explanation methods.

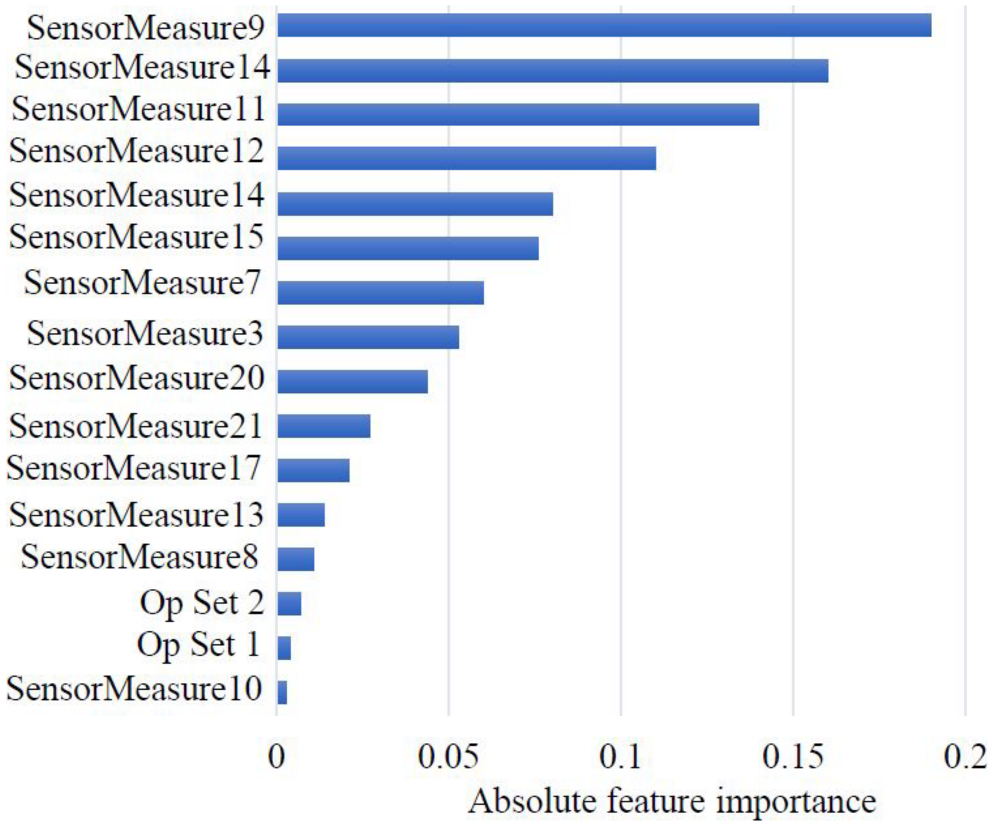



**Figure: For a single prediction, the SHAP-based local explanation**

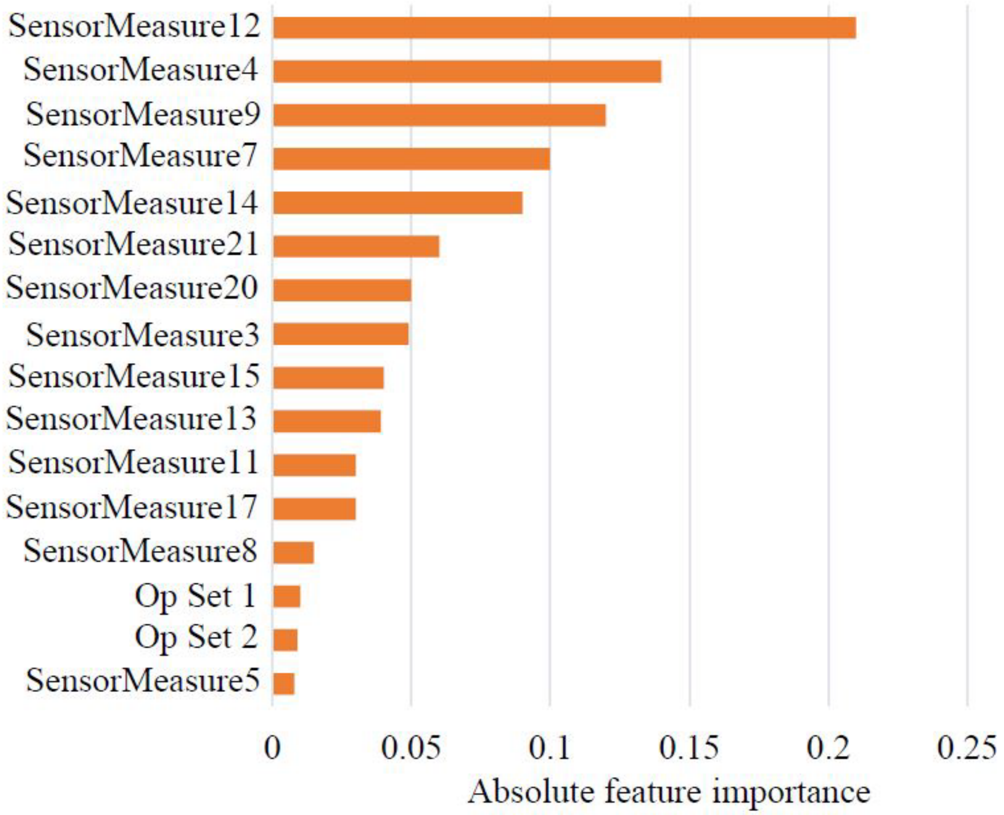



**Figure: For a single prediction, the LIME-based local explanation**

# Proposed approach: Trustworthy RUL explanation

- RUL local explanations method: **SHAP**, **LIME**, and **Anchor**
- Explanation evaluation metrics: **Fidelity**, **Stability**, **Identity**, and **Consistency**
- Ranking and rank aggregation Method: **Kemney** and **Borda** rank aggregation
- **Trust score** measure for **best explanation** method selection

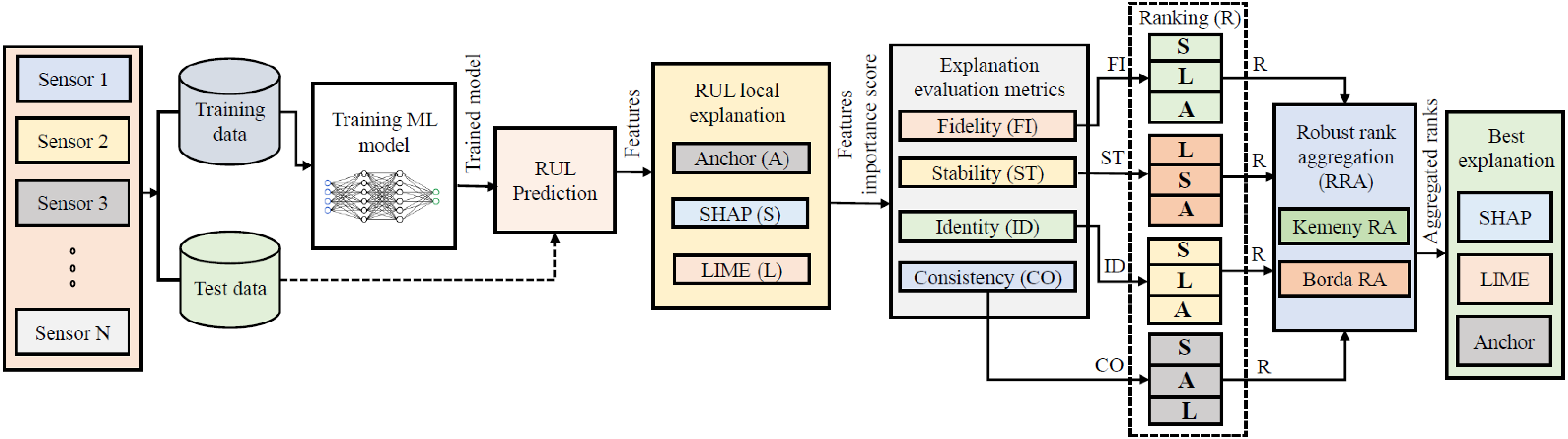



**Figure: An overview of a trustworthy RUL explanation from a set of explanation methods of explainable predictive maintenance framework.**

# RUL local explanation methods

**LIME: Local Interpretable Model-agnostic Explanations**

- Sample points around xi.
- Use a model to predict labels for each sample.
- Weigh samples according to distance to xi.
- Learn simple models on weighted samples.
- Use a simple model to explain.

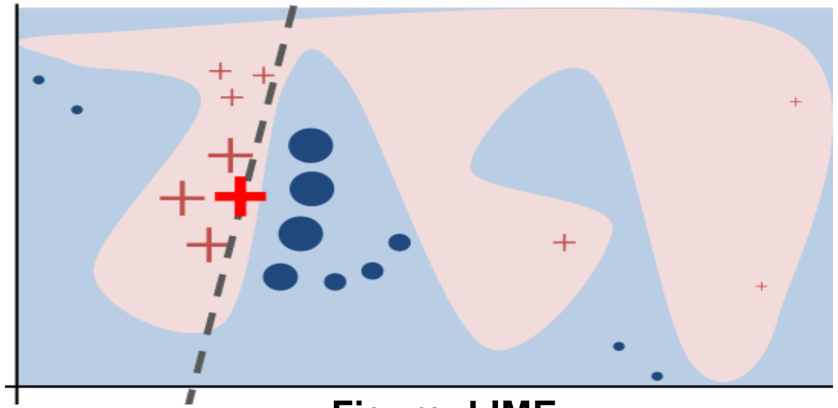


Figure: LIME

**SHAP:SHapley Additive exPlanations**

- **Marginal contribution** of each feature towards the prediction, averaged over all possible permutations.
- **Fairly attributes** the prediction to all the features.

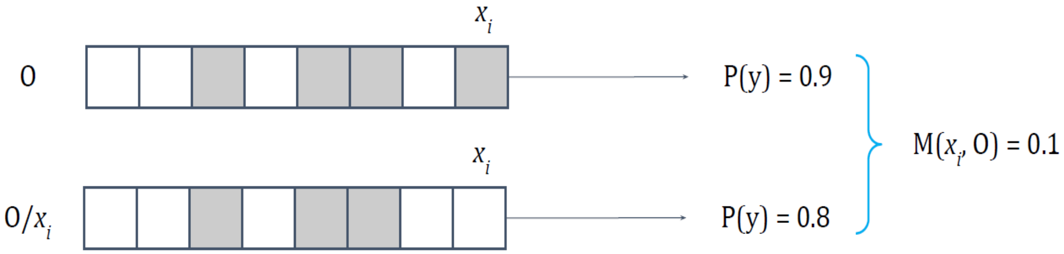



Figure: SHAP

**Anchors**

- Perturb a given instance x to generate a local neighborhood
- Identify an “anchor” rule which has the maximum coverage of the local neighborhood and also achieves a high precision.

**IF** “Operational setting_2” ≥ 0.0034 **AND** “SensorMeasure12”> 522.49
**AND** “SensorMeasure4”≤ 1394.23 **AND** “SensorMeasure9” < 9084.12
**AND** “SensorMeasure14” ≥ 8135.95 **AND** “SensorMeasure7” > 551.60
**AND** “SensorMeasure11” < 48.05 **AND** “SensorMeasure21” ≤ 23.29
**AND** “SensorMeasure15” ≥ 8.38 **AND** “SensorMeasure3” > 1595.65
**THEN PREDICT** “RUL” = 111.87
**WITH** precision = 0.832 **AND** Coverage = 0.232

Figure: Anchors

# Explanation evaluation metrics

**<u>Fidelity</u>**

- To what extent does the explanation method **accurately** represent the underlying decision-making process?
- Explanations that precisely identify the most dominating features of the underlying models for RUL prediction have high fidelity.

**<u>Identity</u>**

- If there are two identical instances, such as the actual and predicted RUL classes, they must have **identical** explanations.
- If this is not the case, then either the explanation model generates an explanation that is **not identical** or the PdM model predicted the wrong RUL class.

**<u>Stability</u>**

- Similar observations should receive **similar** explanations.
- The small changes in the observations will lead to **low changes** in the explanations.

**<u>Consistency</u>**

- Quantifies the **similarity** between the explanations generated by various explanation methods for predictions of different black-box models.
- If an explanation for a single observation is measured multiple times, each of the measured explanations should be **similar**.

# Robust rank aggregation and trust score measure

**Rank aggregation**

- Given a set of rankings $(R_1, R_2, \dots, R_m)$ of a set of objects $(X_1, X_2, \dots, X_n)$ produce a single ranking R that is in agreement with the existing rankings.

**Kemeny**

- Find a barycentric or median ranking by picking a distance on the set of rankings.
- But it is NP-hard to compute.

**Borda**

- For each ranking, assign to object X, a number of points equal to the number of objects it defeats
- The total weight of X is the number of points it accumulates from all rankings

**Trust score (TS)**

- Provides a fair ranking on the performance of aggregated rank and selects the best explanation method for a given predicted RUL.

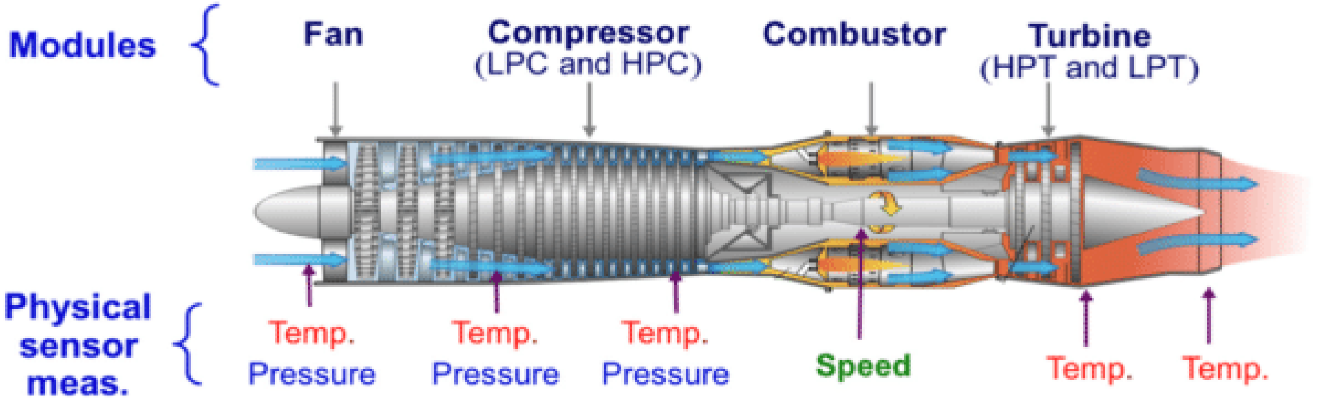
$$\mathrm{TS} = \frac{1}{J}\sum_{p=1}^{N}\sum_{q=1}^{N} Rank_{agr_{score}}(p,q)$$

$Rank_{agr_{score}}(p,q)$ represents the pairwise agreement score between explanation methods p and q in the aggregated rankings and the reference ranking using Kendall's tau ($\tau$) distance.

# Datasets

- Commercial Modular Aero Propulsion System Simulation (C-MAPSS) [1] dataset
  - Pressure
  - Fan speed
  - Fuel
  - Coolant flow
  - Temperature

- Four fleets of engines
  - FD001
  - FD002
  - FD003
  - FD004




**Engine diagram simulated in C-MAPSS [2]**

| | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|
| Train | 100 | 260 | 100 | 249 |
| Test | 100 | 259 | 100 | 248 |
| Op. cond./fault modes | 1/1 | 6/1 | 1/2 | 6/2 |

**Table: Number of train and test engine units in each fleet of the C-MAPSS dataset**

# Results: RUL classification and regression

| Model | MAE | | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|
| | FD001 | FD002 | FD003 | FD004 | FD001 | FD002 | FD003 | FD004 |
| XGB | 13.75 | 15.72 | 14.43 | 18.45 | 14.05 | 16.32 | 14.67 | 17.95 |
| RF | 13.34 | 15.91 | 14.87 | 19.64 | 13.84 | 22.15 | 15.31 | 21.05 |
| LR | 17.55 | 18.71 | 16.23 | 25.87 | 17.76 | 23.03 | 18.32 | 26.92 |
| NN | **9.98** | **11.73** | **10.54** | **12.89** | **12.11** | **14.81** | **13.13** | **14.64** |

Table: Performance of 10-fold cross validation on CMAPSS dataset in RUL prediction task.

| Model | Balanced Accuracy % | | | | F1-Score | | | |
|---|---|---|---|---|---|---|---|---|
| | FD001 | FD002 | FD003 | FD004 | FD001 | FD002 | FD003 | FD004 |
| XGB | 91.5 | 90.3 | 89.7 | 89.3 | 92.6 | 91.4 | 91.2 | 92.5 |
| RF | 89.5 | 88.7 | 88.1 | 87.5 | 91.8 | 90.8 | 91 | 92.1 |
| LR | 87.2 | 86.8 | 84.5 | 85.1 | 90.3 | 89.2 | 88.9 | 89.5 |
| NN | **92.7** | **91.5** | **90.4** | **91.5** | **93.4** | **93.5** | **92.3** | **93.1** |

Table: Performance of 10-fold cross-validation on CMAPSS dataset in the classification task

# Results: SHAP and LIME-based RUL local explanation

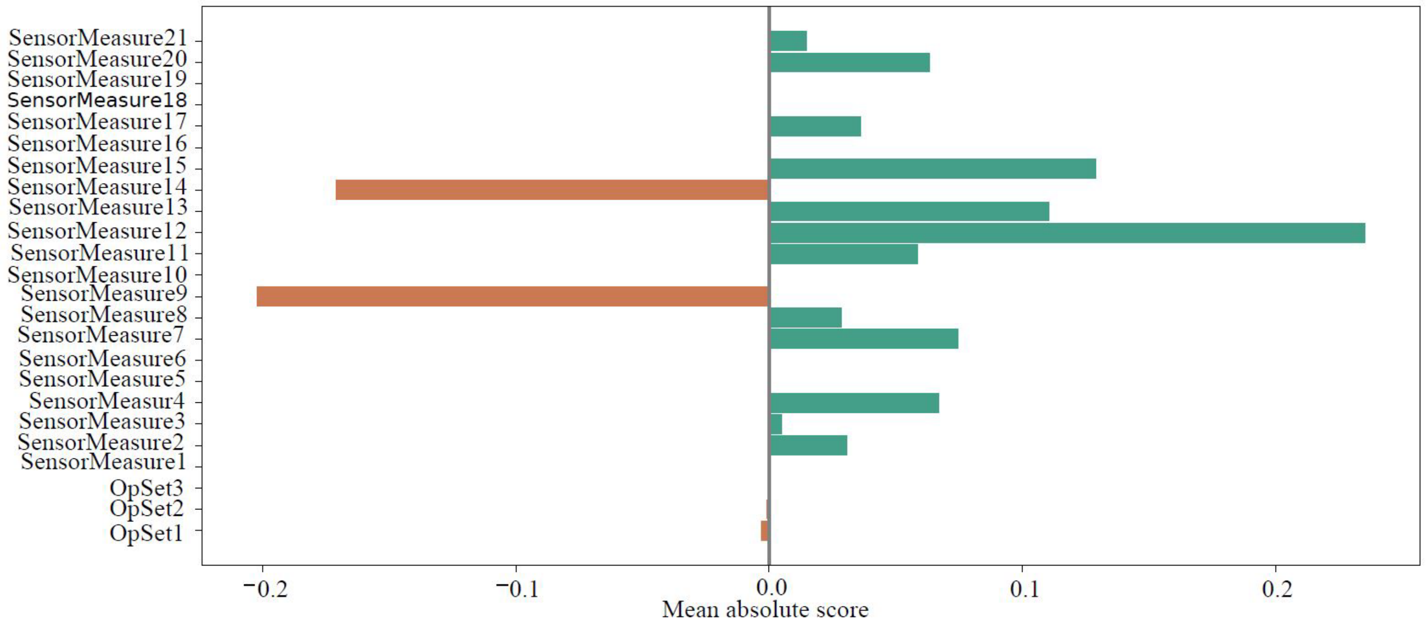



**Figure:** For a single prediction in the FD001 dataset, the local explanations provided by SHAP in which the actual value of RUL of the component is 114 while the predicted value is 111.87 for the FFNN model.

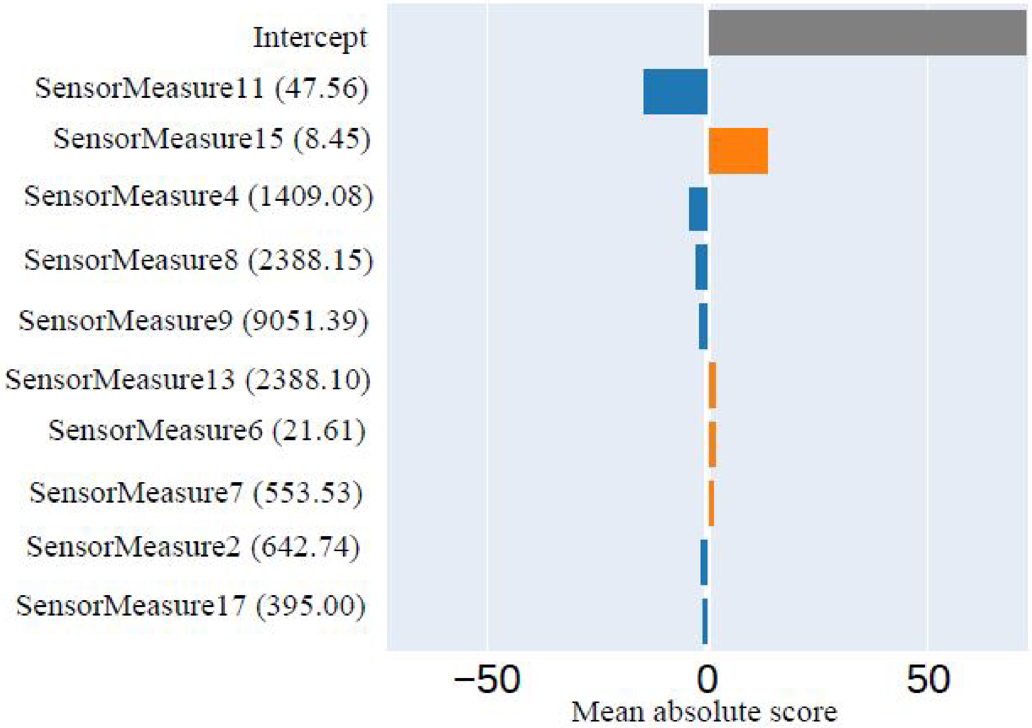



**Figure:** For a single prediction in the FD001 dataset, the local explanations provided by LIME in which the actual value of RUL of the component is 114 while the predicted value is 111.87 for the FFNN model.

# Results: SHAP and Anchor-based RUL local explanation

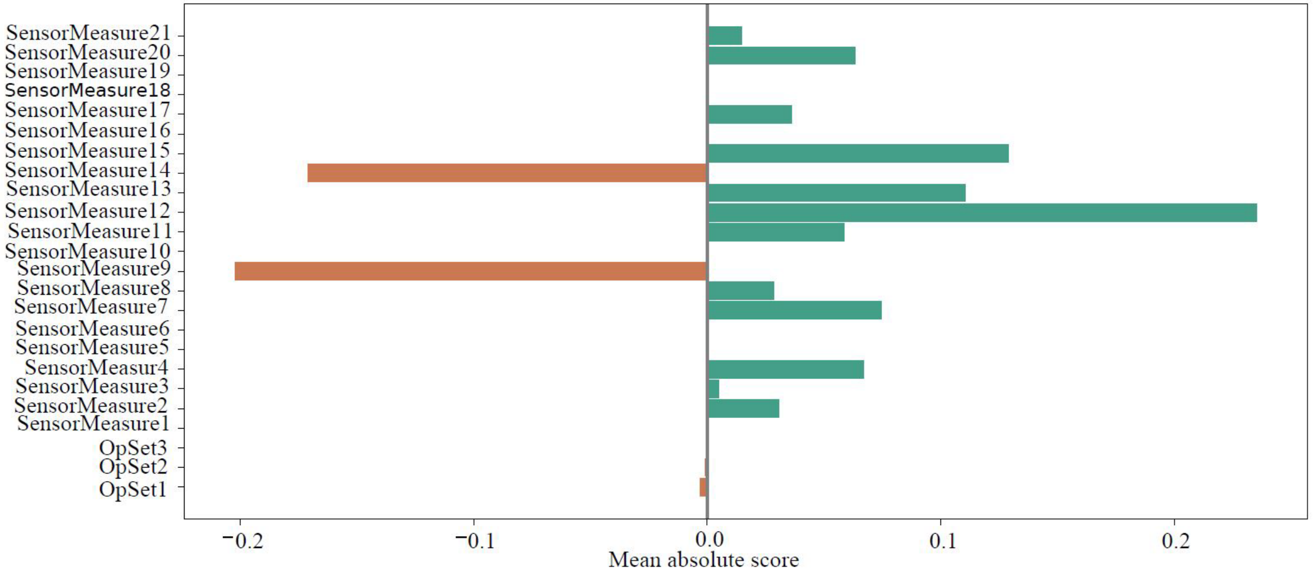



**Figure:** For a single prediction in the FD001 dataset, the local explanations provided by SHAP in which the actual value of RUL of the component is 114 while the predicted value is 111.87 for the FFNN model.

**IF** "Operational setting_2" $\geq$ 0.0034 **AND** "SensorMeasure12"> 522.49
**AND** "SensorMeasure4"$\leq$ 1394.23 **AND** "SensorMeasure9" < 9084.12
**AND** "SensorMeasure14" $\geq$ 8135.95 **AND** "SensorMeasure7" > 551.60
**AND** "SensorMeasure11" < 48.05 **AND** "SensorMeasure21" $\leq$ 23.29
**AND** "SensorMeasure15" $\geq$ 8.38 **AND** "SensorMeasure3" > 1595.65
**THEN PREDICT** "RUL" = 111.87
**WITH** precision = 0.832 **AND** Coverage = 0.232

**Figure:** For a single prediction in the FD001 dataset, the local explanations provided by Anchor in which the actual value of RUL of the component is 114 while the predicted value is 111.87 for the FFNN model.

# Results: Performance of RUL Explanation

| XAI methods | Models | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|---|
| SHAP | LR | 0.875 | 0.843 | 0.795 | 0.892 |
| | XGB | 0.975 | 0.953 | 0.925 | 0.898 |
| | RF | 0.912 | 0.905 | 0.883 | 0.934 |
| | NN | **0.998** | **0.956** | **0.986** | **0.971** |
| LIME | LR | 0.910 | 0.905 | 0.918 | 0.886 |
| | XGB | 0.904 | 0.953 | 0.925 | 0.898 |
| | RF | 0.943 | 0.937 | 0.856 | 0.892 |
| | NN | 0.912 | 0.889 | 0.898 | 0.893 |
| Anchor | LR | 0.863 | 0.843 | 0.795 | 0.892 |
| | XGB | 0.890 | 0.878 | 0.892 | 0.879 |
| | RF | 0.881 | 0.907 | 0.887 | 0.865 |
| | NN | 0.924 | 0.905 | 0.894 | 0.934 |

Table: The fidelity metric of SHAP, LIME, and Anchor methods

| XAI methods | Models | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|---|
| SHAP | LR | 0.032 | 0.0054 | 0.0019 | 0.00056 |
| | XGB | 0.242 | 0.437 | 0.295 | 0.159 |
| | RF | 0.465 | 0.513 | 0.503 | 0.485 |
| | NN | **0.798** | **0.752** | **0.787** | **0.734** |
| LIME | LR | 0.0 | 0.0 | 0.0 | 0.0 |
| | XGB | 0.0242 | 0.0193 | 0.0157 | 0.172 |
| | RF | 0.0805 | 0.081 | 0.061 | 0.074 |
| | NN | 0.08 | 0.053 | 0.079 | 0.071 |
| Anchor | LR | 0.0 | 0.0 | 0.0 | 0.0 |
| | XGB | 0.0 | 0.0 | 0.0 | 0.0 |
| | RF | 0.0 | 0.0 | 0.0 | 0.0 |
| | NN | 0.018 | 0.014 | 0.009 | 0.012 |

Table: The identity metric of SHAP, LIME, and Anchor methods

| XAI methods | Models | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|---|
| SHAP | LR | 0.416 | 0.429 | 0.443 | 0.427 |
| | XGB | 0.339 | 0.353 | 0.331 | 0.319 |
| | RF | 0.302 | 0.325 | 0.336 | 0.317 |
| | NN | **0.273** | **0.295** | **0.301** | **0.289** |
| LIME | LR | 0.507 | 0.537 | 0.525 | 0.519 |
| | XGB | 0.473 | 0.493 | 0.498 | 0.465 |
| | RF | 0.406 | 0.443 | 0.418 | 0.425 |
| | NN | 0.387 | 0.415 | 0.395 | 0.408 |
| Anchor | LR | 0.786 | 0.797 | 0.811 | 0.792 |
| | XGB | 0.687 | 0.703 | 0.719 | 0.749 |
| | RF | 0.745 | 0.762 | 0.716 | 0.704 |
| | NN | 0.642 | 0.669 | 0.638 | 0.655 |

Table: The stability metric of SHAP, LIME, and Anchor methods

| XAI methods | Models | FD001 | FD002 | FD003 | FD004 |
|---|---|---|---|---|---|
| SHAP | LR | 0.0014 | 0.0009 | 0.0008 | 0.001 |
| | XGB | 0.189 | 0.176 | 0.183 | 0.165 |
| | RF | **0.332** | **0.315** | **0.216** | **0.197** |
| | NN | 0.063 | 0.095 | 0.031 | 0.089 |
| LIME | LR | 0.143 | 0.106 | 0.125 | 0.113 |
| | XGB | 0.103 | 0.89 | 0.98 | 0.95 |
| | RF | 0.166 | 0.153 | 0.147 | 0.175 |
| | NN | 0.0087 | 0.059 | 0.0755 | 0.0418 |
| Anchor | LR | 0.0001 | 0.0001 | 0.0001 | 0.0001 |
| | XGB | 0.0032 | 0.0034 | 0.0064 | 0.0009 |
| | RF | 0.0143 | 0.0117 | 0.0122 | 0.0091 |
| | NN | 0.00 | 0.00 | 0.00 | 0.00 |

Table: The consistency metric of SHAP, LIME, and Anchor methods

# Results: Calculating trust scores for identifying the best suitable explanation

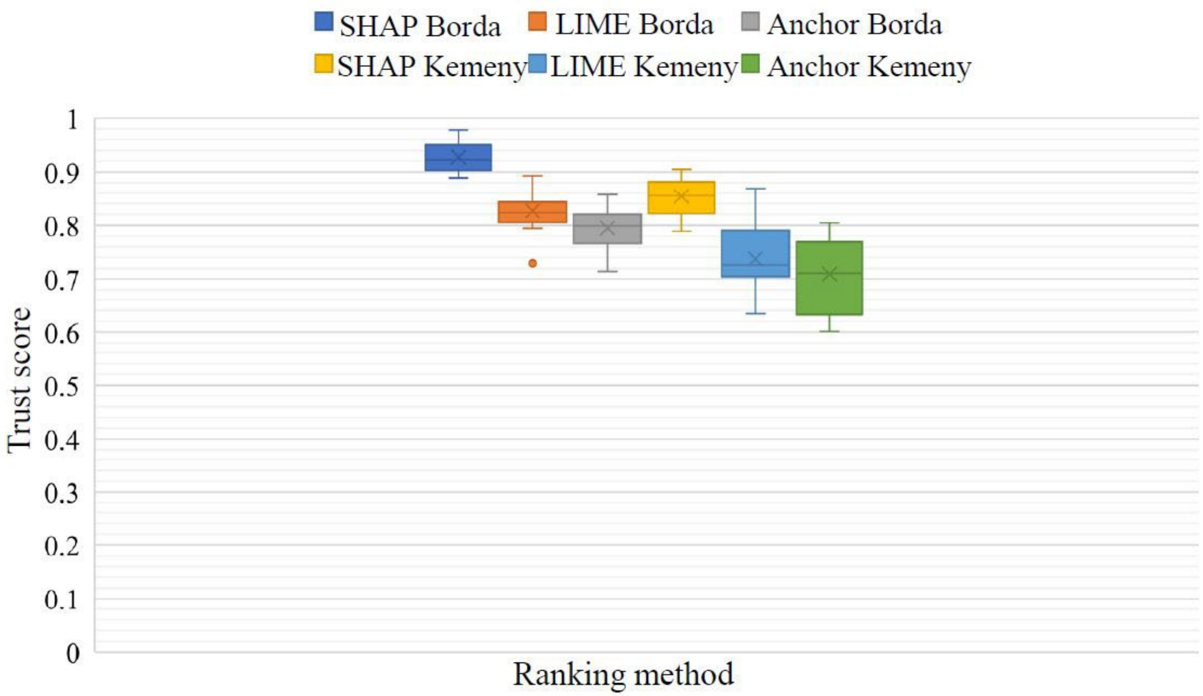



**Figure:** Performance of the top-1 selected model (FFNN-based RUL prediction). Box plots of the measured trust score of the explanation method selected by XAI evaluation metric sets for the FD001 dataset.

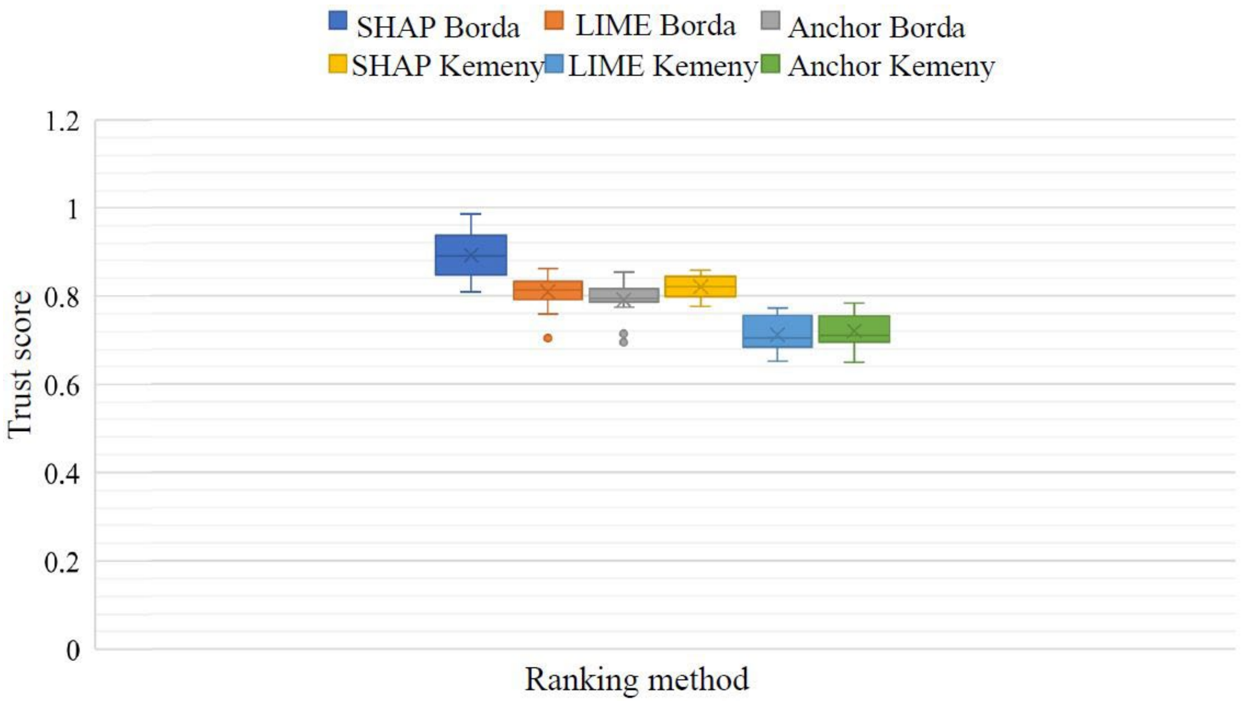



**Figure:** Performance of the top-1 selected model (FFNN-based RUL prediction). Box plots of the measured trust score of the explanation method selected by XAI evaluation metric sets for the FD002 dataset.

# Results: Trustworthy RUL explanation from a set of explanation methods

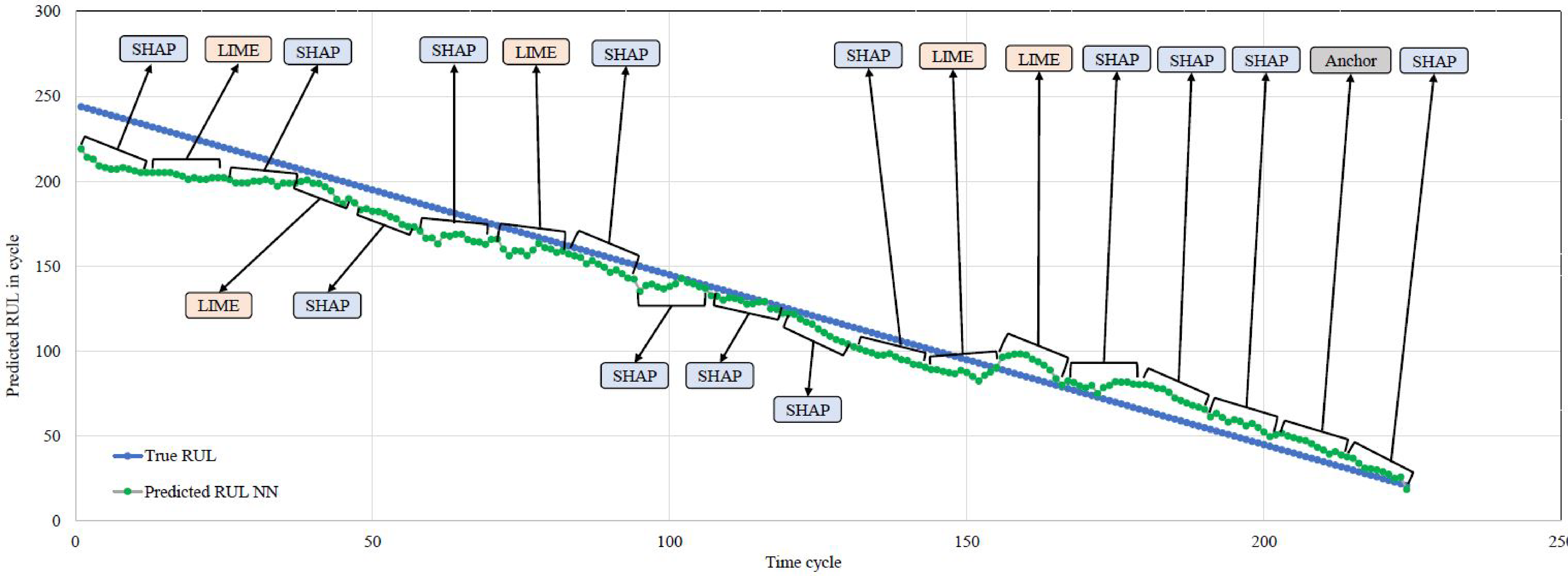



**Figure:** An overview of a trustworthy RUL explanation from a set of explanation methods of explainable predictive maintenance framework using the FFNN model and FD001 dataset.

# Conclusion & Future Work

- Our proposed trustworthy RUL explanation framework by demonstrating and solving the disagreement problem among the state-of-the-art XAI methods.
- Our proposed novel **trust score** by combining their rankings using a robust rank aggregation approach from different explanation evaluation metrics for selecting the best explanation method for a given batch of RUL samples solved the disagreement problem.
- The SHAP explanation method performed relatively well compared to the LIME method.
- The Borda rank aggregation method performed better than the Kemeny method in selecting a suitable explanation method, with the highest **trust score**.
- In future, we plan to conduct further research with other explanation methods such as example-based explanation, counterfactual explanation, visual explanation, etc.

# References


1. Saxena, A., Goebel, K., Simon, D., & Eklund, N. (2008, October). Damage propagation modeling for aircraft engine run-to-failure simulation. In *2008 international conference on prognostics and health management* (pp. 1-9). IEEE.
2. Gungor, O., Rosing, T. S., & Aksanli, B. (2021). Dowell: diversity-induced optimally weighted ensemble learner for predictive maintenance of industrial internet of things devices. *IEEE Internet of Things Journal*, *9*(4), 3125-3134.
3. Jakub Jakubowski, Przemysław Stanisz, Szymon Bobek, and Grzegorz J. Nalepa. 2022. Performance of Explainable AI Methods in Asset Failure Prediction. In Computational Science – ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part IV. Springer-Verlag, Berlin, Heidelberg, 472–485. https://doi.org/10.1007/978-3-031-08760-8_40
4. Khan, T., Ahmad, K., Khan, J., Khan, I., & Ahmad, N. (2022, September). An explainable regression framework for predicting remaining useful life of machines. In *2022 27th International Conference on Automation and Computing (ICAC)* (pp. 1-6). IEEE.
5. Arya, V., Saha, D., Hans, S., Rajasekharan, A., & Tang, T. (2023, January). Global explanations for multivariate time series models. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)* (pp. 149-157).

# Thank you!

## Questions?

**Dependable Cyber-Physical Systems (DCPS) Laboratory**

**Scan me!**