Estimating Cycles to Maintenance Events For Jet Engines Using Engine-specific Measurement Residual

Peihua Han[1], Qin Liang[2]

[1] *Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway*
peihua.han@ntnu.no

[2] *Veracity - DNV, Høvik, Norway*
Qin.Liang@dnv.com

ABSTRACT

This paper introduces a data-driven method for predicting remaining cycles to major maintenance events in commercial jet engines, developed for the PHM North America 2025 Data Challenge. The method leverages measurement residuals that capture sensor deviations from expected values after accounting for operating conditions with simple linear models. These residuals serve as interpretable indicators of engine health. Health indices are constructed for High Pressure Turbine and High Pressure Compressor visits, while Compressor Water Wash events are estimated through linear extrapolation.

1. INTRODUCTION

Predicting maintenance events for jet engines is a critical task in the aviation industry, as timely interventions ensure both operational safety and cost efficiency (Zio, 2022; Liang, Knutsen, Vanem, Æsøy, & Zhang, 2024). With the increasing availability of high-frequency sensor data from commercial aircraft, data-driven methods have emerged as powerful tools for fault detection, anomaly identification, and remaining useful life (RUL) estimation (Han, Ellefsen, Li, Holmeset, & Zhang, 2021; Han, Ellefsen, Li, Æsøy, & Zhang, 2021; Amozegar & Khorasani, 2016; Jiao et al., 2023; Que & Xu, 2019).

Previous research in fault detection and prognostics for complex machinery, including maritime engines and turbines(Han, Li, Skulstad, Skjong, & Zhang, 2020; Liang, Vanem, et al., 2023; Liang, Knutsen, Vanem, Zhang, & Æsøy, 2023; Liang, Vanem, Knutsen, Æsøy, & Zhang, 2024), has demonstrated the effectiveness of using sensor measurements combined with machine learning techniques. Ensemble learning methods and gradient boosting techniques have been widely used for RUL prediction in industrial systems (Jiao et al., 2023; Que & Xu, 2019). In aviation, high-fidelity sensor data captures engine behavior across diverse conditions (Han, Liang, Vanem, Knutsen, & Zhang, 2024), making it difficult to distinguish operating effects from true degradation. Prior work shows that constructing sensor residuals by removing condition influences improves detection of degradation trends (Ellefsen et al., 2020; Vanem et al., 2023; Mathew, Kandukuri, & Omlin, 2024). Using engine-level residuals enables more accurate estimation of remaining cycles to maintenance events, including High Pressure Turbine and High Pressure Compressor shop visits and Compressor Water Wash operations.

The PHM North America 2025 Data Challenge focuses on predicting the cycles of these maintenance events for commercial jet engines. The challenge provides a dataset with engine metadata, sensor readings, and historical maintenance records. The goal is to develop models that accurately estimate the remaining cycles to HPT, HPC, and WW events using engine-level measurements. Our approach uses sensor residuals to construct engine-specific health indicators, which are then used to predict the remaining cycles for critical maintenance events.

2. PROBLEM STATEMENT

The PHM North America 2025 Conference Data Challenge focuses on predicting key maintenance events for commercial jet engines using typically available sensor data. The main objective is to build models to estimate the remaining cycles to three key events:

- High Pressure Turbine (HPT) Shop Visit.
- High Pressure Compressor (HPC) Shop Visit
- HPC Water-Wash (WW)

2.1. Dataset description

The dataset contains metadata and sensor readings from 12 commercial jet engines, each with up to 15,000 data points that span 2,001 flights. As is standard in the commercial aviation industry, each flight provides up to eight snapshots captured at different phases (e.g., takeoff, climb, cruise). A snapshot represents the recorded values of multiple sensors under predefined flight conditions. The dataset is divided by engine: 4 engines for training, 4 for testing, and 4 for validation.

The dataset is organized into three categories: (1) Meta data, (2) Sensor Data, and (3) Targets. A detailed breakdown of the training dataset is provided in Table 1.

Table 1. Details of the training dataset.

Id	Variables
	Meta Data
1	ESN
2	Cycles_Since_New
3	Snapshot
4	Cumulative_WWs
5	Cumulative_HPC_SVs
6	Cumulative_HPT_SVs
	Sensor Data
7	Sensed_Altitude
8	Sensed_Mach
9	Sensed_Pamb
10	Sensed_Pt2
11	Sensed_TAT
12	Sensed_WFuel
13	Sensed_VAFN
14	Sensed_VBV
15	Sensed_Fan_Speed
16	Sensed_Core_Speed
17	Sensed_T25
18	Sensed_T3
19	Sensed_Ps3
20	Sensed_T45
	Targets
21	Cycles_to_WW
22	Cycles_to_HPC_SV
23	Cycles_to_HPT_SV

2.2. Evaluation metrics

To assess prediction performance, a *time-weighted error (TWE)* metric is adopted that penalizes over- and under-predictions asymmetrically and normalizes by the operational horizon of each target. For each prediction–truth pair $(y_i, \hat{y}_i)$, the time-weighted error is defined as

$$\mathrm{TWE}(y_i, \hat{y}_i; \alpha, \beta) = w(y_i, \hat{y}_i) \cdot (\hat{y}_i - y_i)^2 \cdot \beta, \qquad (1)$$

where the weight term is given by

$$w(y_i, \hat{y}_i) = \begin{cases} \dfrac{2}{1 + \alpha y_i}, & \text{if } \hat{y}_i - y_i \geq 0 \\ \dfrac{1}{1 + \alpha y_i}, & \text{if } \hat{y}_i - y_i < 0. \end{cases} \qquad (2)$$

Here, α controls the decay of weights with respect to remaining cycles, and β serves as a normalization factor to ensure comparability across targets with different horizons.

The overall score for each target variable $t \in \{\mathrm{WW}, \mathrm{HPC}, \mathrm{HPT}\}$ is computed as the mean TWE across all samples:

$$\mathrm{Score}_t = \frac{1}{N} \sum_{i=1}^{N} \mathrm{TWE}\big(y_i^{(t)}, \hat{y}_i^{(t)}; \alpha, \beta_t\big), \qquad (3)$$

where N is the number of evaluated snapshots.

Finally, the submission score is obtained by averaging the target-specific scores:

$$\mathrm{Score} = \frac{1}{3} \left(\mathrm{Score}_{WW} + \mathrm{Score}_{HPC} + \mathrm{Score}_{HPT}\right). \qquad (4)$$

This formulation ensures that late predictions (i.e., predicting failures to occur later than they actually do) are penalized more heavily, reflecting the safety-critical nature of maintenance planning in aviation.

3. METHODOLOGY

3.1. Sensor residual construction

Since sensor measurements are influenced not only by the underlying degradation state but also by the operating conditions, it is crucial to separate these two effects. The key idea is to eliminate the influence of operating conditions so that the residual signal more directly reflects degradation. To achieve this, we categorize the sensor measurements into two groups: operating condition–related and degradation-related. We then extract the residual by removing the operating-condition effects using the following formulation:

$$r_d = s_d - f(s_o) \qquad (5)$$

where s_d denotes the degradation-related sensor measurement, s_o is the operating condition–related measurement, and r_d is the resulting residual. *The central assumption is that, under normal conditions, degradation-related sensors can be predicted from the operating-condition sensors through a mapping function f.*

Importantly, residual calculation is performed at the *engine level*: each engine is treated independently and a distinct function f is estimated for each one. Rather than relying on a physics-based model for f, we approximate it with a simple linear regression model, f_{linear}, fitted to the joint data s_o, s_d. Although this assumption is simplified, we found it to be effective in practice. We use $Mach$, $Altitude$, $Pamb$, TAT, $VAFN$, VBV, Fan_Speed, $Pt2$ as s_o and treat the remain-

ing sensors as degradation-related s_d. Using this separation, we compute the residual for each sensor in every snapshot. We observed that the residuals show minimal variation across snapshots, as shown in Figure 1. To obtain overall results, we applied a median filter across the snapshots. The resulting engine-level residuals are presented in Figure 2.

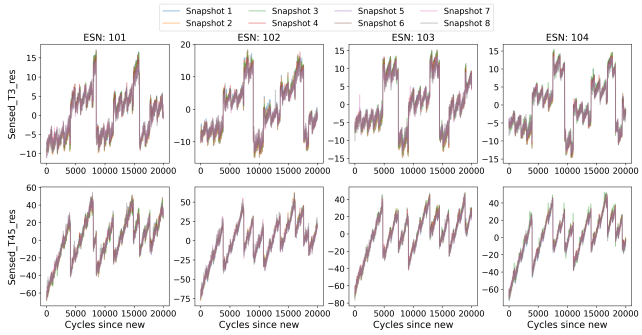


Figure 1. Engine-level sensor residual for T3 and T45 for each snapshot.

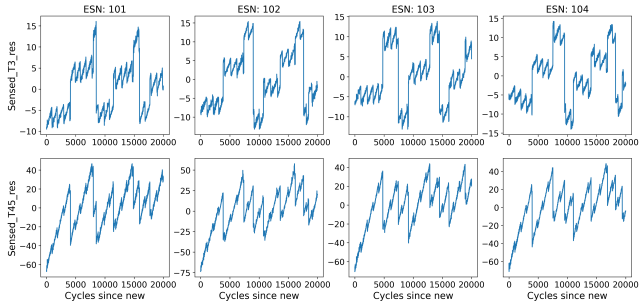


Figure 2. Engine-level sensor residual for T3 and T45 after a median filter.

In Figure 2, it is evident, particularly for $T3_{res}$ and $T45_{res}$, that the HPT, HPC, and WW events are clearly reflected as sudden jumps in the residual signals.

3.2. Cycles to HPT estimation

To estimate cycles to HPT, we construct the HPT health index HI_{HPT} as a linear combination of $T3_{res}$ and $T45_{res}$:

$$HI_{HPT} = -\alpha_{HPT} T3_{res} - T45_{res} \tag{6}$$

where α_{HPT} is an engine-specific coefficient determined by minimizing the deviation from the HPC reference.

Figure 3 shows the joint plot of HI_{HPT} with $Cycles\_to\_HPT$. The results suggest that, for most engines, $Cycles\_to\_HPT$ can be approximated as a linear function of HI_{HPT}, except at high cumulative HPT service for some engines. It should be noted that the mapping between HI_{HPT} and $Cycles\_to\_HPT$ is also engine-specific.

For the test and validation sets, determining the engine-specific α_{HPT} and linear mapping from HI_{HPT} to $Cycles\_to\_HPT$ is not straightforward due to the short measurement windows. To address this, we first group the short-window samples by engine number. Within each short window, HPC, HPT, and WW events can be clearly identified from $T3_{res}$. We then use the HPC event to determine α_{HPT} for each engine, and use the HPT event to establish the linear mapping between HI_{HPT} and $Cycles\_to\_HPT$ by assuming that the cycles equal zero at the occurrence of the HPT event.

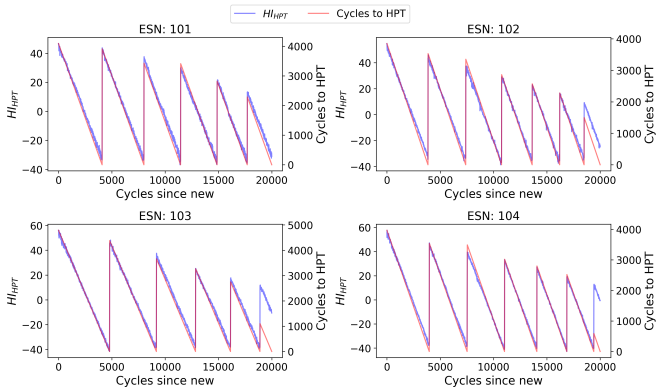


Figure 3. Joint plot of HI_{HPT} and $Cycles\_to\_HPT$.

3.3. Cycles to HPC estimation

Similarly, the HPC health index HI_{HPC} is formulated as a linear combination of $T3_{res}$ and $T45_{res}$:

$$HI_{HPC} = -\alpha_{HPC} T3_{res} - T45_{res} \tag{7}$$

where α_{HPC} is an engine-specific coefficient determined by minimizing the deviation from the HPT reference.

Figure 4 shows the joint plot of HI_{HPC} with $Cycles\_to\_HPT$. The results suggest that HPC events cannot be separated from WW events. When the cumulative HPC service reaches 2, some deviation appears between HI_{HPT} and $Cycles\_to\_HPT$. Nevertheless, we continue to apply an engine-specific linear mapping between HI_{HPC} and $Cycles\_to\_HPC$.

Similarly for the test and validation sets, we first group the short-window samples by engine number and identify HPC, HPT, and WW events from $T3_{res}$ in each short window. We then use the HPT event to determine α_{HPC} for each engine, and use the HPC event to establish the linear mapping between HI_{HPC} and $Cycles\_to\_HPC$ by assuming that the cycles equal zero at the occurrence of the HPC event.

When the cumulative HPC service reaches 2, we build a LightGBM model to classify it and then quantify the gap between the true service cycle and our linear prediction. We find that this gap is linearly correlated with both the slope

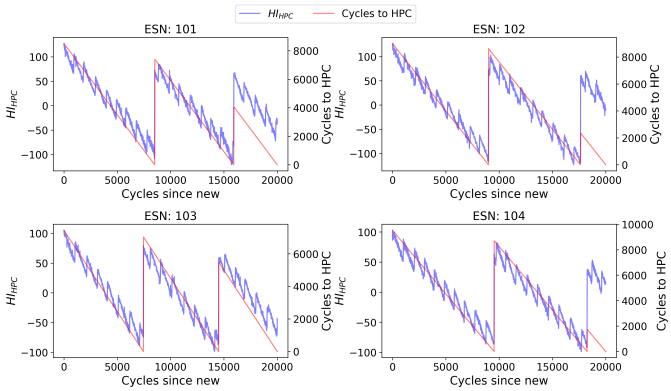


Figure 4. Joint plot of HI_{HPC} and $Cycles\_to\_HPC$.

and the intercept of the linear mapping from HI_{HPC} to $Cycles\_to\_HPC$, as illustrated in Figure 5.

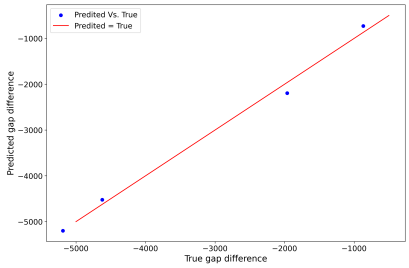


Figure 5. Gap predicted by the slope and the intercept of the linear mapping from HI_{HPC} to $Cycles\_to\_HPC$.

For the test and validation sets, we use the LightGBM model to classify whether the cumulative HPC service reaches 2. If at least 30% of the short-window predictions are positive, we apply a gap correction derived from the slope and intercept.

3.4. Cycles to WW estimation

To estimate the cycles to WW, we rely solely on $T45_{res}$. A closer examination of $T45_{res}$ reveals that, once the last WW event is identified, the timing of the next WW event is governed by two factors: (1) the slope of $T45_{res}$increase, and (2) the increment of $T45_{res}$ until the next WW event. Analysis of the training dataset shows that both factors can be reasonably approximated as constants.

Figure 6 shows the $T45_{res}$ after removing the effects of HPC, HPT, and WW events. It can be observed that the residuals follow an approximately linear trend, which can be captured by the fitted curve. The slope of the fitted curve is 0.029, indicating that $T45_{res}$ increases by about 2.9 every 100 cycles in the absence of HPC, HPT, and WW events.

Figure 6 shows the $T45_{res}$ after removing the effects of HPC and HPT events. We then extracted the values of $T45_{res}$ at each WW event. The results indicate that $T45_{res}$ at the WW event points also increases approximately linearly with

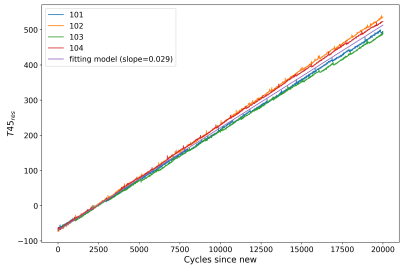


Figure 6. $T45_{res}$ after removing the effects of HPC, HPT, and WW events.

the number of WW events. The slope of the fitted curve is 0.029, suggesting that $T45_{res}$ increases by approximately 21 per WW event in the absence of HPC and HPT events.

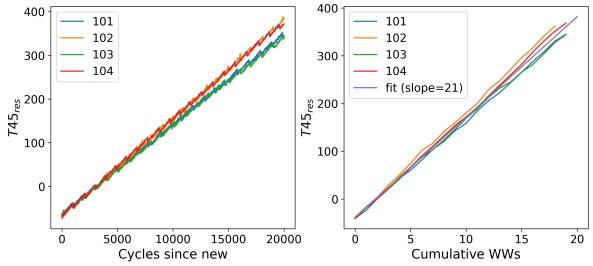


Figure 7. Left: $T45_{res}$ after removing the effects of HPC and HPT. Right: $T45_{res}$ at the WW event points versus cummulative WWs.

4. RESULTS AND DISCUSSIONS

4.1. Performance

Table 2 and Table 3 summarize the test and final evaluation results, respectively. Our method ranked 4th in the test set but dropped to 18th in the final validation set. Notably, this discrepancy is not unique to our approach: all of the top five teams in the test set experienced a significant drop in ranking on the validation set, with large score differences. This suggests that there may be substantial differences between the test and validation sets.

Table 2. Test Result.

Rank	Team Name	Score
#1	MathWorks	0.3528
#2	WISDOM	1.802
#3	ICDI	13.57
#4	**PHHQ**	**21.17**
#5	aeae	23.22
#6	lookhill	36.28
#7	SAM-IPA-1	37.11
#8	Justin_Boredom	37.22

From our perspective, we believe that our approach provides reasonably accurate estimates for HPT and WW events in

Table 3. Final Validation Result.

Rank	Team Name	Score
#1	SAM-IPA-1	47.54
#2	lookhill	48.56
#3	Justin_Boredom	49.3
...	...	...
#10	aeae	88.99
#12	WISDOM	96.46
#13	MathWorks	97.69
#14	SAM-IPA-1	37.11
#15	ICDI	104.1
...	...	...
#18	**PHHQ**	**128.7**

both the test and validation sets, but is less accurate for HPC events. The scoring system penalizes over-predictions and places greater importance on cases with low remaining cycles. Since cycles to HPC can reach up to 12,500, even a modest over-prediction can result in a large penalty. For example, if the true HPC is 500 cycles and the prediction is 1,500, the score for that sample would be 1,333; if the prediction is 2,500, the score jumps to 5,333. Given that there are only 47 samples in the test and validation sets, a single outlier can disproportionately affect the overall score.

4.2. Examples

Figure 8 and Figure 9 show examples of applying our approach to estimate cycles to HPT and HPC, respectively. The light blue line represents the original estimation, while the dark blue line shows the estimation after correcting for HPT or HPC events. The red line represents our final estimation, with its endpoint on the y-axis corresponding to our predicted result.

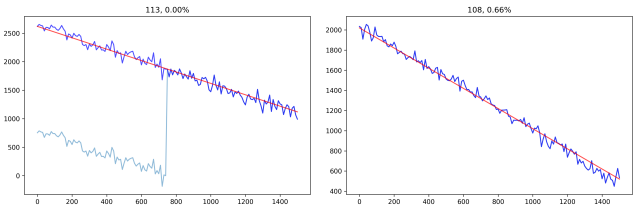


Figure 8. Examples of estimating cycles to HPT in the validation set.

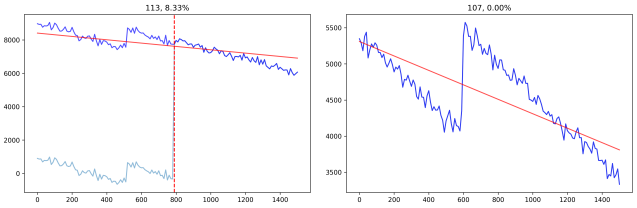


Figure 9. Examples of estimating cycles to HPC in the validation set.

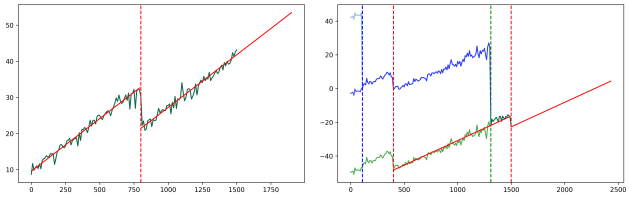


Figure 10. Examples of estimating cycles to WW in the validation set.

Figure 10 shows examples of applying our approach to estimate cycles to WW. The green line represents the signal after correcting for both HPT and HPC events. The red vertical lines indicate the WW events. The red line, which follows the trend of the green signal, represents our estimation, with its endpoint on the x-axis corresponding to the predicted WW cycle.

5. CONCLUSION

This paper presents a solution for estimating the cycles of jet engine maintenance events in the PHM North America 2025 Conference Data Challenge. The core idea is to construct engine-specific sensor residuals that capture the degradation state while removing the influence of operating conditions.

REFERENCES

Amozegar, M., & Khorasani, K. (2016). An ensemble of dynamic neural network identifiers for fault detection and isolation of gas turbine engines. *Neural Networks*, *76*, 106–121.

Ellefsen, A. L., Han, P., Cheng, X., Holmeset, F. T., Æsøy, V., & Zhang, H. (2020). Online fault detection in autonomous ferries: Using fault-type independent spectral anomaly detection. *IEEE Transactions on instrumentation and measurement*, *69*(10), 8216–8225.

Han, P., Ellefsen, A. L., Li, G., Æsøy, V., & Zhang, H. (2021). Fault prognostics using lstm networks: application to marine diesel engine. *IEEE Sensors Journal*, *21*(22), 25986–25994.

Han, P., Ellefsen, A. L., Li, G., Holmeset, F. T., & Zhang, H. (2021). Fault detection with lstm-based variational autoencoder for maritime components. *IEEE Sensors Journal*, *21*(19), 21903–21912.

Han, P., Li, G., Skulstad, R., Skjong, S., & Zhang, H. (2020). A deep learning approach to detect and isolate thruster failures for dynamically positioned vessels using motion data. *IEEE Transactions on Instrumentation and Measurement*, *70*, 1–11.

Han, P., Liang, Q., Vanem, E., Knutsen, K. E., & Zhang, H. (2024). Assessing helicopter turbine engine health: A simple yet robust probabilistic approach. In *Annual conference of the phm society* (Vol. 16).

Jiao, Z., Wang, H., Xing, J., Yang, Q., Yang, M., Zhou, Y., & Zhao, J. (2023). Lightgbm-based framework for lithium-ion battery remaining useful life prediction under driving conditions. *IEEE Transactions on Industrial Informatics*, *19*(11), 11353–11362.

Liang, Q., Knutsen, K. E., Vanem, E., Æsøy, V., & Zhang, H. (2024). A review of maritime equipment prognostics health management from a classification society perspective. *Ocean Engineering*, *301*, 117619.

Liang, Q., Knutsen, K. E., Vanem, E., Zhang, H., & Æsøy, V. (2023). Unsupervised anomaly detection in marine diesel engines using transformer neural networks and residual analysis. In *Phm society asia-pacific conference* (Vol. 4).

Liang, Q., Vanem, E., Knutsen, K. E., Æsøy, V., & Zhang, H. (2024). Anomaly detection in time series data: A novel approach using transformer neural networks for reconstruction and residual analysis. *International Journal of Prognostics and Health Management*, *15*(3).

Liang, Q., Vanem, E., Xue, Y., Alnes, Ø., Zhang, H., Lam, J., & Bruvik, K. (2023). Data-driven state of health monitoring for maritime battery systems–a case study on sensor data from ships in operation. *Ships and Offshore Structures*, 1–13.

Mathew, M. S., Kandukuri, S. T., & Omlin, C. W. (2024). Soft ordering 1-d cnn to estimate the capacity factor of windfarms for identifying the age-related performance degradation. In *Phm society european conference* (Vol. 8, pp. 9–9).

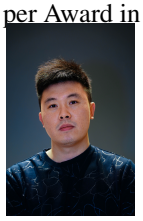
Que, Z., & Xu, Z. (2019). A data-driven health prognostics approach for steam turbines based on xgboost and dtw. *IEEE Access*, *7*, 93131–93138.

Vanem, E., Liang, Q., Ferreira, C., Agrell, C., Karandikar, N., Wang, S., ... others (2023). Data-driven approaches to diagnostics and state of health monitoring of maritime battery systems. In *Proceedings of the annual conference of the phm society 2023*.

Zio, E. (2022). Prognostics and health management (phm): Where are we and where do we (need to) go in theory and practice. *Reliability Engineering & System Safety*, *218*, 108119.

BIOGRAPHIES

Peihua Han received the Ph.D. degree in engineering from the Norwegian University of Science and Technology (NTNU), Aalesund, Norway, in 2022, where he is now a senior researcher. His Ph.D. thesis was recognized with the CHOROFAS Prize in 2022. His research interests include data mining, machine learning, time series modeling, and uncertainty quantification. He has published about 50 papers in peer-reviewed journals and conferences and received the IEEE Robotics and Automation Magazine Best Paper Award in 2024.



Qin Liang works as the Maritime Data Architect at Veracity, DNV, where he leads the development of data architecture strategies for maritime emissions and decarbonization solutions. He previously served as a Senior Researcher in Group Research and Development at DNV (2018–2025) and as a Data Scientist in Ship Intelligence at Rolls-Royce Marine (2015–2018). He received his Ph.D. in Engineering from the Norwegian University of Science and Technology (NTNU) in 2025, his M.Sc. in Product and System Design from NTNU in 2015, and his B.Sc. in Marine Engineering from Dalian Maritime University in 2013. His research interests include ship performance optimization, equipment condition monitoring, machine learning, deep learning, and maritime battery health management.