Assessing Helicopter Turbine Engine Health: A Simple Yet Robust Probabilistic Approach

Peihua Han[1], Qin Liang[2], Erik Vanem [3], Knut Erik Knutsen [4], Houxiang Zhang[5]

[1,2,5] *Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway*
peihua.han@ntnu.no
qinlia@stud.ntnu.no
hozh@ntnu.no

[2,3,4] *Group Research and Development - DNV, Høvik, Norway*
Qin.Liang@dnv.com
Erik.Vanem@dnv.com
Knut.Erik.Knutsen@dnv.com

Abstract

This paper presents a data-driven approach for assessing the health of helicopter turbine engines, developed for the PHM North America 2024 Conference Data Challenge. The task involves both regression and classification to estimate the torque margin and classify engine health as either nominal or faulty. To quantify the reliability of predictions, probabilistic outputs are generated. We employ a two-stage model where the predicted torque margin serves as an input feature for health classification. For probabilistic torque margin estimation, we introduce an empirical error sampling method to generate torque margin samples, followed by a rule-based distribution selection scheme to evaluate the resulting distributions. For fault classification, logistic regression is used to provide confidence estimates, and we incorporate a score-optimized loss function during training to mitigate penalties for false negatives. Our approach demonstrates strong generalization to unseen assets, ranking 2nd in the competition with a score of 0.94, demonstrating its effectiveness in predicting health conditions and uncertainty for more informed helicopter engine management.

1. Introduction

A turbine engine, also known as a gas turbine or jet engine, is an internal combustion engine that turns fuel into mechanical energy. In helicopters, these engines are crucial for providing the power needed for lift and maneuvering. They operate under high-stress conditions and experience significant wear and tear, which can lead to failures. Effective health assessment is essential to predict potential problems, prevent major failures, and lower maintenance costs. It can ensure that engines operate safely and efficiently, supporting reliable helicopter performance.

Prognostics and Health Management (PHM) is an integrated framework designed to monitor, diagnose, and predict the condition of systems (Zhang et al., 2022). It has been applied to many applications including turbine engines. PHM encompasses three key components: anomaly detection (Han, Ellefsen, Li, Holmeset, & Zhang, 2021), fault diagnostics (Wang et al., 2020), and fault prognostics (Han, Ellefsen, Li, Æsøy, & Zhang, 2021). PHM methods can be categorized into model-based and data-driven approaches depending on whether a physical model is used. Data-driven methods have gained popularity in PHM due to their ability to handle complex, high-dimensional data and uncover patterns that may be challenging for traditional model-based approaches to capture (Liang, Knutsen, Vanem, Æsøy, & Zhang, 2024).

In data-driven settings, fault diagnostics can be addressed through regression or classification. When using regression, the target output typically represents the system's degradation level, modeled as a continuous variable. For example, Vanem et al. (2023); Liang, Vanem, et al. (2023) estimated the state of health of a battery by extracting features from charging and discharging curves and applying various statistical models. Similarly, Mathew et al. (2024) developed a one-dimensional convolutional neural network to estimate the capacity factor of wind farms, utilizing the Huber loss function to mitigate the impact of outliers. In classification tasks, the target output is a categorical variable, e.g., fault detection often involves binary outputs, while fault isolation deals with multiclass outputs that represent specific isolation outcomes. Amozegar and Khorasani (2016) proposed an ensemble approach combining three different machine learning models: multi-layer perceptron, radial basis function neural network, and support vector machine, to detect and isolate faults in gas turbine engines. Han et al. (2020) employed a one-dimensional convolutional neural network with focal loss to handle imbalanced datasets when isolating thruster faults in marine vessels. A review of statistical methods for condition monitoring is presented in (Vanem, 2018); see also e.g. (Hastie, Tibshirani, & Friedman, 2009) for a thorough introduction to such methods.

The difference between classical statistical modelling and machine learning has long been discussed in the data science community (Breiman, 2001; Carmichael & Marron, 2018; Bzdok, Altman, & Krzywinski, 2018). Although it may be debated whether there is a fundamental difference (Tibshirani & Hastie, 2021), the distinction between the models capability of explaining and predicting have been raised (Shmueli, 2010), and explainability has emerged as an important topic, in particular for complicated ”black box” models (Adadi & Berrada, 2018; Kakavandi, Han, De Reus, Larsen, & Zhang, 2023; Liang, Knutsen, Vanem, Zhang, & Æsøy, 2023). Another issue with more complicated models is generalizability and the risk of overfitting to a data sample (Gohil et al., 2024). Although neural network based models are known to be universal approximators that are able to describe very complicated patterns in a dataset, simpler models may be more robust in estimating the underlying structure in the population. Hence, the principle of parsimony is an argument to choose a simpler model, with fewer parameters, over a more complicated model if it performs equally well. Notwithstanding, the regression and classification models presented in this paper are relatively simple models, but they are found to perform very well on the data challenge problem at hand.

In addition, probabilistic outputs with uncertainty estimates are particularly valuable for PHM (Zio, 2022), as they allow better maintenance scheduling and more effective risk management. For practical deployment, it is crucial to quantify the uncertainty and confidence in detection, diagnostics, and prognostics, enabling more informed decision-making regarding the operation of engineering systems. One potential approach is the use of probabilistic models, such as Gaussian processes, Bayesian neural networks, or ensemble methods like bootstrap aggregation, to effectively capture uncertainty.

In this paper, we outline the development of a data-driven model for assessing the health of helicopter turbine engines as part of the PHM North America 2024 Conference Data Challenge. Our approach generates probabilistic outputs that predict the performance distribution of the engine and provide confidence estimates regarding whether the engine is classified as faulty or nominal.

2. PROBLEM STATEMENT

The PHM North America 2024 Conference Data Challenge focuses on evaluating the health of helicopter turbine engines, addressing both regression and classification tasks. The objective is to assess engine health by predicting a key variable known as the torque margin, which quantifies the extent of engine underperformance, and by classifying whether the engine is faulty. Additionally, the challenge requires reporting the confidence levels for both the regression and classification outputs, making it a probabilistic regression and classification problem.

To summarize, for each observation in the dataset, the task is to predict the asset's health by:

- Estimating the torque margin as a probability distribution function (PDF).
- Predicting the binary health state (0 = nominal, 1 = faulty), along with a confidence metric represented as a continuous variable ranging from 0 to 1.

2.1. Dataset description

The combined dataset consists of seven helicopter engines (assets), all of the assets are the same make and model. Measured data for four of these assets is provided in the training set, though the observations have been shuffled and asset IDs removed to anonymize the data. The remaining three assets are allocated to the test and validation sets.

Each engine is equipped with a range of sensors that capture various measurements, as detailed in Table 1. The dataset is structured as a tabular format, comprising 742,624 observations in the training set, 21,436 observations in the test set, and 21,436 observations in the validation set.

Table 1. Sensor measurements from the dataset.

Variables	**Abbreviation**	**Type**
outside air temperature	oat	float
mean gas temperature	mgt	float
pressure altitude	pa	float
indicated airspeed	ias	float
net power	np	float
compressor speed	ng	float
torque measured	$T_{measured}$	float

For each operational condition, there is a specified design (target) torque, which represents the expected performance of the engine. The actual output torque is also measured. Engine health is evaluated by comparing the measured output torque to the target torque. Specifically, the torque margin, an indicator of engine health, is calculated as:

$$T_{margin} = 100 \times \frac{(T_{measured} - T_{target})}{T_{target}} \tag{1}$$

where T_{margin} represents the torque margin, $T_{measured}$ is the measured torque, T_{target} and is the target torque.

The objective is to estimate the torque margin T_{margin} and determine the health state (0 = nominal, 1 = faulty) using the sensor measurements provided in Table 1.

2.2. Evaluation metrics

A key objective of this competition is to assess the confidence associated with the submitted classification and regression predictions. Confidence levels play a crucial role in the scoring system. The classification and regression performance will be evaluated independently, and the final score will be the average of the classification and regression scores across all predictions.

Classification score: The classification score will be linearly weighted for correct predictions and false positives, with a strong penalty for highly confident false negatives. This is because a false negative, which occurs when the model predicts the engine is healthy when it is actually faulty, can lead to costly repairs and, in the worst-case scenario, serious safety risks. Figure 2 illustrates the regression score across all possible scenarios.

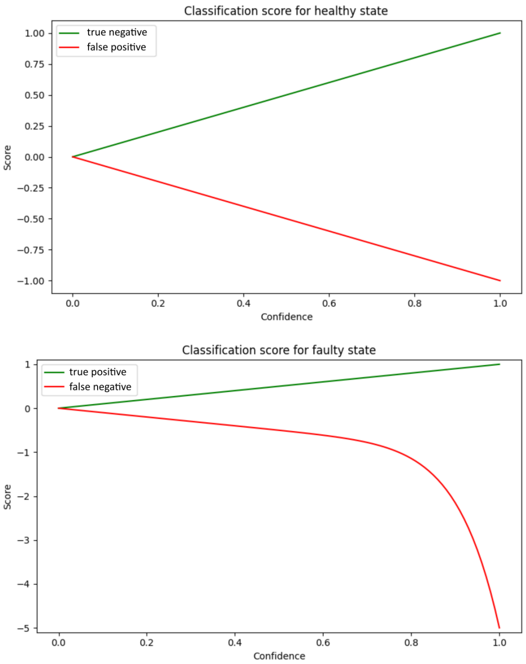

Figure 1. An illustration of the classification score regarding healthy and faulty states from official website (PHMSociety, 2024).

Regression score: The regression score is computed based on the intersection of the true value with the predicted probability density function (PDF). If the maximum value of the PDF exceeds 1, the score is normalized by the maximum PDF value to prevent excessively high scores in cases where the PDF is very narrow. Figure 2 illustrates examples of predictions and their corresponding scores using Normal and Cauchy PDFs.

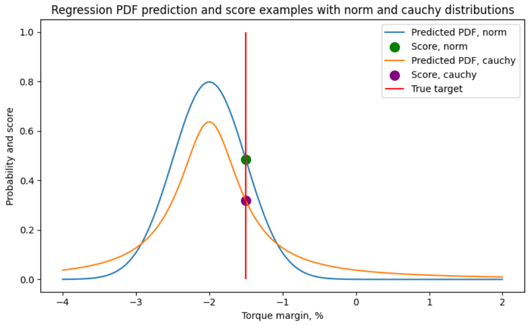

Figure 2. An illustration of the regression score from official website (PHMSociety, 2024).

3. METHODOLOGY

3.1. Overview of the method and feature engineering

Instead of treating regression and classification as independent tasks, we adopt a two-stage approach. First, the torque margin is predicted, and this predicted torque margin is then used as a feature for classification. Figure 3 provides an overview of the proposed method.

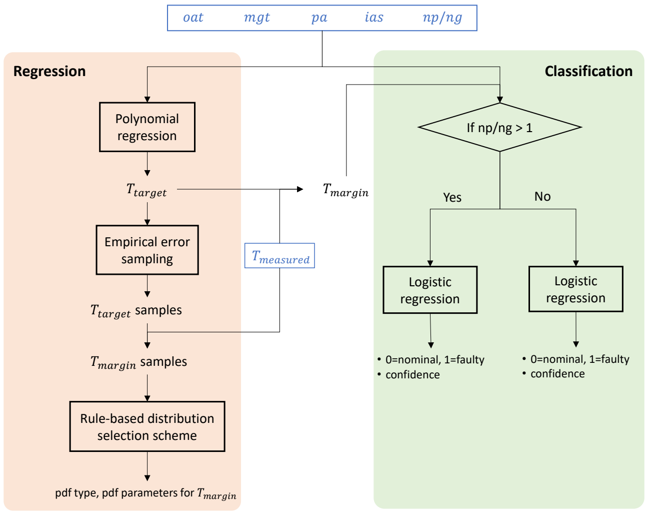

Figure 3. Overview of the methodology.

Instead of using net power np and compressor speed ng directly, we introduce a new feature, np/ng. The np/ng ratio acts as a key indicator of how efficiently the engine converts compressor speed into power. A higher ratio suggests that the engine is producing more power for a given compressor speed, indicating improved efficiency. Conversely, a lower ratio may signal inefficiencies or a mismatch between speed and power output. By incorporating this feature, our model effectively captures the subtle dynamics between compressor speed and power generation. Note that a standard normalization is performed for all the features.

For the regression task, instead of predicting the torque margin T_{margin} directly, we predict the torque target T_{target} and compute the torque margin using the measured torque values $T_{measured}$. This approach is motivated by the observation that features such as mgt and np/ng exhibit strong linear correlations with T_{target}. In the classification task, two logistic regression models are developed based on whether np/ng is less than 1.

Figure 4 illustrates the relationship between mgt, np/ng, and T_{target}. Both mgt and np/ng show strong linear correlations with T_{target}, which led us to predict the torque target T_{target} rather than the torque margin T_{margin}. Additionally, there is a clear distinction based on whether the np/ng ratio is greater or less than 1. When $np/ng < 1$, the nominal and faulty cases become linearly separable. This observation supports our decision to adopt a two-model classification approach.

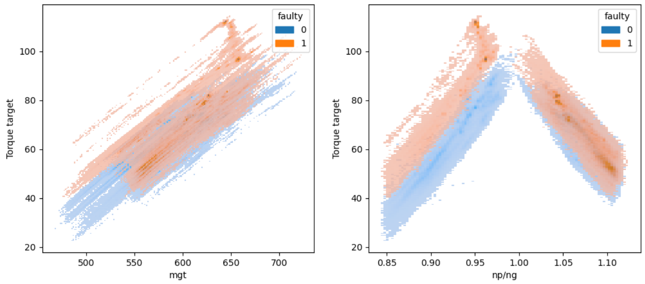

Figure 4. mgt, np/ng ratio with torque target

Key components of the methodology, such as empirical error sampling, the rule-based distribution selection scheme, and the score-optimized loss in logistic regression, are detailed in the following sections.

3.2. Empirical error sampling for probabilistic output

In order to predict the torque target, a polynomial regression model is employed, utilizing the input features: mgt, oat, ias, pa, np/ng. The model is defined as:

$$\hat{y} = \beta^\top \Phi(\mathbf{x}) \tag{2}$$

where $\Phi(\mathbf{x})$ represents the polynomial transformation of the input feature vector $\mathbf{x}$, incorporating all polynomial combinations of the features up to the third order (as selected in this study). Hence, the model can be regarded as a type of multivariable fractional polynomial regression (Royston & Altman, 1994; Sauerbrei & Royston, 1999). The vector β contains the corresponding regression coefficients, and the intercept term is absorbed into β. Notably, the polynomial features include interaction terms up to the third order, such as x_1x_2 and $x_1^2 \cdot x_2$. An example of how the model calculates third-order features is provided below:

$$\begin{aligned} \hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{n_1} x_1^2 + \cdots \\ + \beta_{n_2} x_1 \cdot x_2 + \beta_{n_3} x_2^3 + \cdots + \beta_{n_4} x_1^2 \cdot x_2 + \cdots \end{aligned} \tag{3}$$

The error term for each training sample is calculated as the residual between the observed target output y_i and the predicted output $\hat{y}_i$:

$$\epsilon_i = y_i - \hat{y}_i = y_i - \beta^\top \Phi(\mathbf{x}_i) \tag{4}$$

The complete set of training residuals is stored in the set $\mathcal{E} = \epsilon_1, \epsilon_2, \ldots, \epsilon_n$, where n represents the total number of training samples.

To introduce a probabilistic aspect into the model's predictions, we perform error sampling from the empirical distribution of the training residuals. For a new input feature vector $\mathbf{x}_{\text{new}}$, a sample from the probabilistic prediction $\hat{y}_{\text{new}}^{(j)}$ is obtained by adding a randomly sampled residual $\epsilon^{(j)}$ from $\mathcal{E}$ to the predicted output:

$$\hat{y}_{\text{new}}^{(j)} = \beta^\top \Phi(\mathbf{x}_{\text{new}}) + \epsilon^{(j)}, \quad \epsilon^{(j)} \sim \mathcal{E} \tag{5}$$

By repeating this sampling process m times (with $m = 1000$ in this study), a set of probabilistic predictions is generated $\{\hat{y}_{\text{new}}^{(1)}, \hat{y}_{\text{new}}^{(2)}, \ldots, \hat{y}_{\text{new}}^{(m)}\}$. This ensemble of predictions reflects the distribution of possible outcomes, allowing for a probabilistic interpretation of the model's output.

3.3. Rule-based distribution selection scheme

From the previous section, we have generated a set of samples for the torque target, denoted as $\{\hat{T}_{target}^{(1)}, \hat{T}_{target}^{(2)}, \ldots, \hat{T}_{target}^{(m)}\}$, to represent the probabilistic output distribution. However, transforming this torque target sample set into a torque margin sample set $\{\hat{T}_{margin}^{(1)}, \hat{T}_{margin}^{(2)}, \ldots, \hat{T}_{margin}^{(m)}\}$ is nontrivial, as the measured torque values are available.

Rather than fitting a statistical distribution to the sample set, we define four distinct distributions, detailed in Table 2, to approximate the sample set while aiming to maximize the score. The design of these distributions is to ensure that the maximum probability density function value is 1.

An illustration of the four designed distributions is provided in Figure 5. The selection process follows a predefined rule: distributions are evaluated in order of priority. If 99% of the samples fall within a given distribution, that distribution is selected. If none of the first three distributions satisfy this rule, the Cauchy distribution is applied. This design ensures that no overconfident predictions will be given, as the regression score normalizes when the PDF exceeds 1.

Table 2. Designed distributions.

Order	Distribution	Parameters
1	Uniform	$loc = \hat{T}_{margin} - 0.5, scale = 1$
2	Beta	$a = 1.5,\ b = 1.5,$ $loc = \hat{T}_{margin} - 0.6365,\ scale = 1.273$
3	Normal	$loc = \hat{T}_{margin}, scale = 1/\sqrt{2\pi}$
4	Cauchy	$loc = \hat{T}_{margin}, scale = 1/\pi$

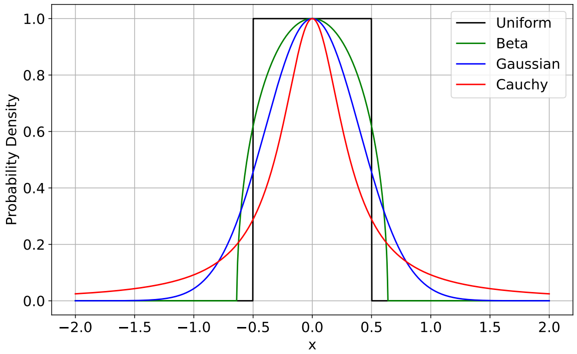

Figure 5. An illustration of the 4 designed distribution.

3.4. Score-optimized loss in logistic regression

In the binary classification task, logistic regression is employed to predict the probability of class 1 (faulty), see e.g. (Hastie et al., 2009). The standard logistic regression model outputs the probability as:

$$\hat{y} = \sigma(\mathbf{w}^\top \mathbf{x}) \tag{6}$$

where $\mathbf{x}$ is the feature vector. $\mathbf{w}$ is the vector of model weights. σ is the sigmoid function, which maps the output to the range [0,1].

Typically, the log-likelihood is used as the loss function to estimate the weights $\mathbf{w}$. The log-likelihood loss is known for producing well-calibrated models, where the predicted probabilities accurately reflect the confidence of the predictions. However, in this study, we introduce a custom loss function that penalizes false negatives more heavily. This adjustment reflects the critical nature of misclassifying faulty instances, where a well-calibrated model may not necessarily lead to the best classification score.

To address this, we design a custom loss function that directly optimizes the classification score in Figure. 1. The custom loss function is defined as:

$$\begin{aligned}\mathcal{L}_{custom} =& (1-y)(1-\hat{y}) - (1-y)\hat{y} + y\hat{y} \\ & - y\left(4(1-\hat{y})^{11} + (1-\hat{y})\right)\end{aligned} \tag{7}$$

where y is the ground truth binary class and $\hat{y}$ is the predicted probability of class 1. The custom loss function is designed to penalize false negatives more heavily, particularly through the last term, which includes a regression score raised to the power of 11. The objective of this custom loss function is to maximize the classification score, shifting the focus from producing well-calibrated probabilities to improving performance in terms of specific classification scores.

4. RESULTS AND DISCUSSIONS

4.1. Performance

Table 3 summarizes the final evaluation results for the top 10 competition entries. Our approach achieved 2nd place with a score of 0.94, highlighting the effectiveness of our methodology.

Table 3. Final Evaluation Result.

Rank	Team Name	Score
#1	goldriver	0.996590
#2	**PHHQ**	**0.940693**
#3	ajouPHM	0.917999
#4	MathWorks	0.913785
#5	Sliding Kurtosis Rules!	0.904316
#6	ppgeps	0.886953
#7	Ajoucau	0.866422
#8	Mad SoftMax	0.848976
#9	SuperNOVa	0.840498
#10	B-26418	0.787819

Table 4 presents the results of an ablation study on our approach, conducted on both the training and test sets. It's important to note that the test scores are automatically generated after submission, and we only have access to the overall score, preventing us from separately evaluating the regression and classification components. We experimented with polynomial features of varying orders for both tasks, as well as with and without a custom loss function for classification. For the regression task, higher-order polynomial features led to better results, with a particularly notable improvement when using second-order features compared to linear models. However, in classification, while the increased polynomial order improved the training score, it caused a significant drop in the test score, indicating overfitting. The use of a custom loss function yielded a modest improvement in the training score but resulted in a substantial gain in the test score.

4.2. Feature importance

One advantage of the adopted model comparing with other complex and advanced models is the explainability. It is straightforward to interpret how the model calculates the fi-

Table 4. Scores on train and test set with different model settings.

Model	Train			Test
	Regr	Cls	Total	Total
Final model	**0.999**	0.867	0.933	**0.984**
w 2-order poly (regr)	0.994	0.867	0.930	0.983
w/o poly (regr)	0.748	0.867	0.808	0.857
w/o custom loss (cls)	**0.999**	0.867	0.933	0.975
w 3-order poly (cls)	**0.999**	**0.953**	**0.976**	0.936
w 2-order poly (cls)	**0.999**	0.893	0.946	0.938

nal output based on the trained coefficient values. All features are standard normalized prior to training the regression and classification models. Figure 6 presents the feature importance of the top 15 features in the regression task. The most significant features, such as mgt, oat, and pa, dominate the regression model output. These features are physically meaningful: Higher mgt leads to greater expansion of gases, which increases torque. Lower oat and lower pa improve air density, allowing for more efficient combustion, which also boosts torque. These relationships demonstrate the strong dependence of engine torque on both thermal conditions (mgt) and environmental factors (oat, pa).

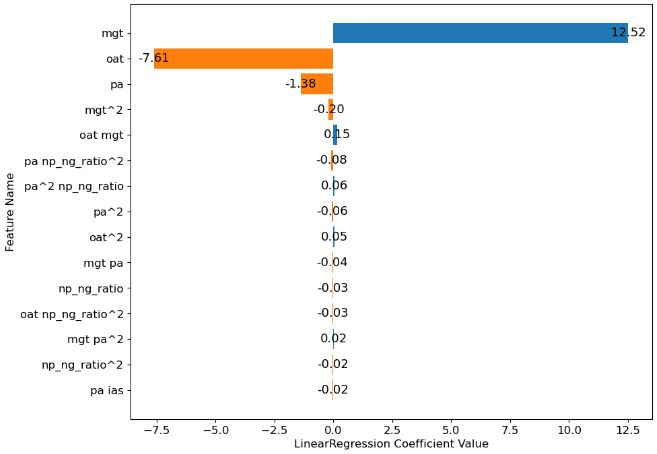

Figure 6. Feature importance of the regression model.

Figure 7 shows the feature importance for the classification model. The most influential feature is the torque margin, with a large negative coefficient, indicating that a lower torque margin significantly increases the likelihood that the engine is classified as faulty. The torque margin represents the difference between the measured and target torque, making it a direct indicator of engine underperformance.

4.3. Discussions

The results show that our approach achieves near-perfect scores for regression but performs less optimally in classification. Throughout the competition, we experimented with incorporating higher-order and custom features. While this significantly improved the training score, the test score declined, suggesting potential overfitting. This highlights the need for further enhancing classification performance in future iterations.

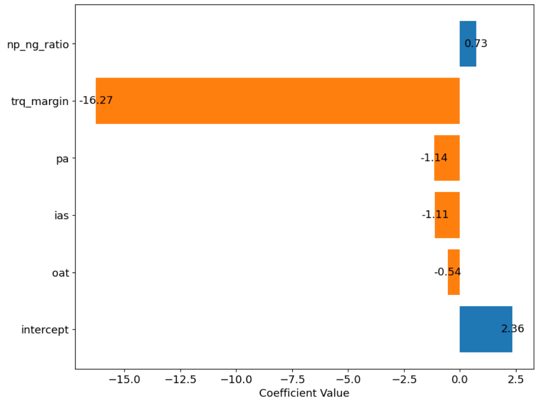

Figure 7. Feature importance of the classification model.

Additionally, several other advanced models have been tested, including LightGBM (Jiao et al., 2023), Random Forest (Liang, Vanem, Knutsen, & Zhang, 2022), XGBoost (Que & Xu, 2019), and MLP (Multi-Layer Perceptron) (Liang, Tvete, & Brinks, 2019, 2020). These strong models can all get a perfect score on the training set, but it is less satisfatory when evaluated on test data in daily assessments. And perform a train-val spilt on training set did not help, mainly contributes that the engine in training set are shuffle, making it impossile to test the generalization performance of the model. In addtion, the MLP yields a much better performance than tree-based models, this could potentially due to the reason that tree-based models are not able to generate a continous smooth decision boundary.

The scoring function for regression normalizes the probability density function to 1, which inadvertently penalizes highly confident predictions. This motivated us to design a rule-based distribution selection scheme in our approach. However, this creates a dilemma: the model can exploit this by using a uniform distribution for all predictions, diminishing the importance of capturing uncertainty in the process.

5. CONCLUSION

This paper presents a winning solution for assessing helicopter turbine engine health in the PHM North America 2024 Conference Data Challenge. The success of our approach is driven by three key principles:

- Simplicity enhances generalization: We employed low-order polynomial and linear models, which proved effective across unseen data.
- Torque target prediction: By predicting torque target instead of torque margin, we simplify the problem.

- Thoughtful probabilistic design: We carefully constructed probabilistic output distributions to avoid overconfident predictions, ensuring accurate regression scoring.

Our model is designed to handle both regression and classification tasks using sensor measurements from helicopter turbine engines. It generates probabilistic outputs, providing not only predictions but also insights into the confidence of those predictions. While the regression component achieved near-perfect scores, there is room for improvement in the classification aspect, indicating that future work could focus on refining the classification performance.

ACKNOWLEDGMENTS

This work was carried out as part of the RealTOPs project (grant number 331634), supported by DNV and the Research Council of Norway. The authors declare no competing interests

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138 - 52160.

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.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199-231.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233-234.

Carmichael, I., & Marron, J. (2018). Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, *1*, 117-138.

Gohil, V., Dev, S., Upasani, G., Lo, D., Ranganathan, P., & Delimitrou, C. (2024). The importance of generalizability in machine learning for systems. *IEEE Computer Architecture Letters*, *23*(1), 95-98.

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

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.

Kakavandi, F., Han, P., De Reus, R., Larsen, P. G., & Zhang, H. (2023). Interpretable fault detection approach with deep neural networks to industrial applications. In *2023 international conference on control, automation and diagnosis (iccad)* (pp. 1–7).

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., Tvete, H., & Brinks, H. (2020). Prediction of vessel propulsion power from machine learning models based on synchronized ais-, ship performance measurements and ecmwf weather data. In *Iop conference series: Materials science and engineering* (Vol. 929, p. 012012).

Liang, Q., Tvete, H. A., & Brinks, H. W. (2019). Prediction of vessel propulsion power using machine learning on ais data, ship performance measurements and weather data. In *Journal of physics: Conference series* (Vol. 1357, p. 012038).

Liang, Q., Vanem, E., Knutsen, K. E., & Zhang, H. (2022). Data-driven prediction of ship propulsion power using spark parallel random forest on comprehensive ship operation data. In *2022 ieee 17th international conference on control & automation (icca)* (pp. 303–308).

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).

PHMSociety. (2024). PHM North America 2024 Conference Data Challenge. *https://data.phmsociety.org/phm2024-conference-data-challenge/*.

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

Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *43*(2), 429-452.

Sauerbrei, W., & Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(1), 71-94.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289-310.

Tibshirani, R., & Hastie, T. (2021). A melting pot. *Observational Studies*, *7*(1), 213-215.

Vanem, E. (2018). Statistical methods for condition monitoring systems. *International Journal of Condition Monitoring*, *8*, 9-23.

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.*

Wang, T., Li, G., Skulstad, R., Æsøy, V., & Zhang, H. (2020). An effective model-based thruster failure detection method for dynamically positioned ships. In *2020 ieee international conference on mechatronics and automation (icma)* (pp. 898–904).

Zhang, H., Li, G., Hatledal, L. I., Chu, Y., Ellefsen, A., Han, P., . . . Hildre, H. P. (2022). A digital twin of the research vessel gunnerus for lifecycle services: Outlining key technologies. *IEEE Robotics & Automation Magazine*, *30*(3), 6–19.

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 bachelor's in civil engineering and master's degree in structural engineering from the Department of Architecture and Civil Engineering, Zhejiang University, Hangzhou, China, in 2016 and 2019, respectively, and the Ph.D. degree in engineering from the Norwegian University of Science and Technology (NTNU), Aalesund, Norway, in 2022. He is currently a senior researcher at NTNU. His current research interests include data mining, machine learning, time-series modeling, and uncertainty qualification.

Qin Liang works as a Senior Researcher in Group Research and Development - Maritime programe DNV. He worked as a Data Scientist in Ship Intelligence with Rolls Royce Marine (2015-2018). In additon, he is currently pursuing the Ph.D. degree with the Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology (NTNU). His current research interests include ship performance, equipment condition monitoring, machine learning and deep learning.

Erik Vanem received the Cand. Scient. degree (Master of Science equivalent) in physics and the Ph.D. degree in statistics from the University of Oslo in 1996 and 2012, respectively. He worked three years at the Research Department of Telenor, three years at PGS Reservoir, one year at the Oslo University College, and spent some time at the Norwegian Defence Research Establishment. He has been working at legacy DNV R&I since 2003 on a number of research projects related to maritime safety and risk assessment. Since 2016, he has also been an Associate Professor with the University of Oslo in a 20% position. He is currently working as a Principal Researcher with the Maritime Transport Group, DNV Group Research and Development, Høvik, Norway. As a Researcher, he has authored and coauthored a number of papers in international journals and international conference proceedings and authored a recent monograph.

Knut Erik Knutsen is a Principal Researcher at DNV – Group Research and Development – Maritime Transport and team lead for the Data Driven Services group. He holds a PhD in Semiconductor Physics (2013) and a MSc in Materials, Energy and Nanotechnology (2009) from the University of Oslo. He also has experience as an Avionics Technician from the Royal Norwegian Airforce (1999-2004) where he was performing maintenance and calibration of aircraft equipment. Currently his research focuses on Data Driven Services and in particular data integrity solutions on the edge to increase the level of trust in data driven applications, and further enable novel digital solutions to support DNVs purpose of safeguarding life, property and the environment.

Houxiang Zhang received the Ph.D. degree in mechanical and electronic engineering and the Habilitation degree in informatics from the University of Hamburg, Hamburg, Germany, in 2003 and 2011, respectively. Since 2004, he has been a Postdoctoral Fellow and a Senior Researcher with the Institute of Technical Aspects of Multimodal Systems, Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, University of Hamburg, Germany. In April 2011, he joined the Norwegian University of Science and Technology (NTNU), Aalesund, Norway, where he is currently a Professor of mechatronics. From 2011 to 2016, he also held a Norwegian National GIFT Professorship on product and system design funded by the Norwegian Maritime Centre of Expertise. He has applied for and coordinated more than 30 projects supported by the Norwegian Research Council (NFR), German Research Council (DFG), EU, and industry. In these areas, he has authored or co-authored more than 250 journals and conference papers as the author or coauthor. He has engaged in two main research areas, including control, optimization, and AI application, especially on autonomous vehicles, marine automation, digitalization, and ship intelligence.