Robust Health Condition Prediction of Helicopter Turboshaft Engines Using Ensemble Machine Learning Models

Zihan Wu, Junzhe Wang, and Meng Li

NOV, Houston, Texas, 77042, United States of America

zihan.wu@nov.com
eric.wang@nov.com
meng.li@nov.com

ABSTRACT

This paper presents a novel ensemble approach that combines multiple machine-learning algorithms to deliver robust predictions of helicopter turboshaft engine health status (nominal or faulty) using operational data. Engine health is evaluated through the torque margin, defined as the percentage difference between the measured and target torque values. A Gaussian process model is used to estimate the torque margin as a probability distribution function (PDF), and these predictions are incorporated as features into various machine-learning models. These models are then employed to perform binary classification, determining the engine's health state. To enhance performance, a reference set is defined for each unseen data point, allowing a comparison of the relative performances of the models, with the best performer selected for the final prediction. Our ensemble method achieves high accuracy in health classification while providing precise torque margin estimates. The results demonstrate that ensemble models offer superior generalization and reliability compared to individual machine-learning algorithms, especially when applied to complex, multivariate datasets like those from helicopter turboshaft engines.

1. INTRODUCTION

Helicopter turboshaft engines are complex mechanical systems whose health is critical to the safety and performance of aviation operations. Effective monitoring and prediction of engine health can prevent costly failures and ensure operational readiness (Elasha et al., 2021). Traditional maintenance practices rely heavily on scheduled inspections, which can lead to unnecessary downtime or missed detection of early-stage faults (Achouch et al., 2022, Wu et al., 2023). To address these limitations, there is a growing interest in data-driven approaches (Daouayry et al., 2018,) that leverage operational data to assess engine health in real time and enable predictive maintenance strategies. A comprehensive review of data-driven prognostic methods was provided in (Schwabacher et al., 2005). These approaches typically involve the fusion of sensor data, feature extraction, and statistical pattern recognition. For predicting the health condition, techniques such as interpolation (Wang et al., 2008), extrapolation (Coble et al., 2008), or machine learning (Wu et al., 2022) are often employed, among others.

Machine learning (ML) techniques have shown great potential in condition monitoring (Surucu et al., 2023) and fault detection (Nelson et al., 2023, Wang et al., 2024, Zheng et al., 2024) across various industries. In particular, ensemble methods—where multiple machine learning models are combined to improve prediction accuracy and robustness—have proven effective in handling complex, multivariate datasets. By leveraging diverse models, ensemble methods can capture different patterns and relationships within the data, leading to improved generalization on unseen assets (Mian et al., 2024).

In this work, we propose an ensemble machine-learning framework for predicting the health of helicopter turboshaft engines using operational measurements such as outside air temperature, compressor speed, and torque margin. Our approach focuses on two key tasks: (1) binary classification of engine health state (nominal or faulty) and (2) probabilistic regression to estimate the torque margin. This dual-task framework provides not only a fault diagnosis but also a confidence metric for the torque margin, enhancing the interpretability of the predictions.

The contributions of this paper are twofold: first, we introduce an ensemble learning model for engine health classification and torque margin estimation; second, we validate our approach on a dataset of seven engines, demonstrating its generalization capability across unseen assets. Our results show that the proposed ensemble method outperforms individual machine learning models, offering a reliable solution for predictive maintenance in helicopter turboshaft engines.

2. Exploratory data analysis

Before applying advanced machine learning techniques, it is crucial to perform thorough Exploratory Data Analysis (EDA) to verify the quality of the data, uncover underlying patterns, and address any inconsistencies. In this study, a key aspect of EDA involves analyzing engine performance data by examining seven critical measured parameters as well as one two important feature engineering parameters. In the provided data, all turboshaft engines are equipped with sensors that capture these seven parameters: outside air temperature (OAT), mean gas temperature (MGT), available power (PA), indicated airspeed (IAS), net power (NP), compressor speed (NG), and engine torque (Trq_measured).

Outside Air Temperature (OAT) plays a role in engine performance by directly influencing the engine inlet air temperature, which in turn affects the thermodynamic balance within the engine and may have an impact on overall engine health.

Mean Gas Temperature (MGT) serves as a key indicator of engine health by representing the thermal state within the engine. Changes in MGT often reflect underlying issues such as shifts in combustion quality, which may result from inefficiencies or increased engine stress. As such, monitoring MGT is essential for assessing whether the engine is operating under optimal conditions, with deviations potentially indicating reduced fuel efficiency or mechanical wear.

Available Power (PA) reflects the maximum potential output of the engine under current operating conditions. While it doesn't directly cause changes in performance, discrepancies between expected and available power can indicate issues such as degraded engine components or suboptimal operating conditions, suggesting potential health concerns.

Indicated Airspeed (IAS) represents the aircraft's velocity, indirectly influencing engine workload. At certain airspeeds, the engine failure rate may be higher, requiring special attention to these conditions to ensure engine reliability and safety.

Net Power (NP), which is the effective power output after accounting for system losses, serves as an indicator of overall engine efficiency. A significant reduction in net power compared to expected values may signal mechanical issues or inefficiencies within the engine system, such as frictional losses or degraded components, impacting engine health.

Compressor Speed (NG) reflects the rotational speed of the compressor, which regulates airflow and pressure within the engine.

Torque represents the rotational force produced by the engine that drives the main rotor. It is a key parameter for assessing engine performance, as it directly correlates to the engine's ability to produce the necessary power to lift and maneuver the helicopter.

In addition to directly measured sensor parameters, feature engineering is applied. Using the six previous operational parameters, a target torque is derived based on empirical correlations, representing the designed torque. The percentage difference between the designed and measured torque, referred to as the torque margin, is calculated using Equation 1 and serves as a key indicator of engine performance and health. In this study, the health of turboshaft engines is classified into two states: healthy or faulty, depending on whether the torque margin deviates from expected values.

$$Trq_{margin} = 100 * \left(\frac{Trq_{measured} - Trq_{target}}{Trq_{target}}\right) \quad (1)$$

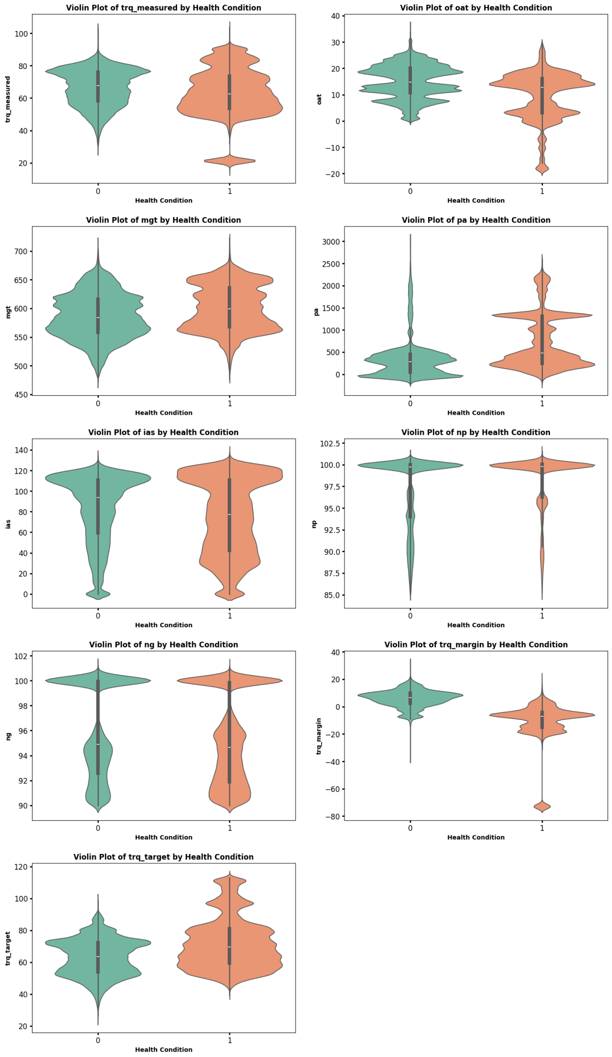

Figure 1 Violin plot of parameters by health conditions

In this study, we utilized operational data and torque margin measurements from four identical turboshaft engines, each assessed under varying health conditions. The dataset comprises 742,625 operational data points, including parameters such as engine performance and torque margins under different health states. The target torque value can be calculated using Equation 1. To avoid sequence bias, the data were shuffled. Notice that there are 443,207 healthy data points and 29,9418 faulty data points within the training dataset, indicating that the data imbalance is not significant. However, this does not necessarily mean that the data is balanced in specific regions, particularly in the testing data region. Further investigation is needed.

Figure 1 presents violin plots illustrating the distribution characteristics of various parameters for helicopter engines under healthy (0) and faulty (1) conditions in the training dataset. It is evident that different torque margin and torque target ranges are associated with varying engine failure rates. Notably, when the torque margin falls below a certain threshold or the torque target exceeds a specified limit, it consistently indicates a 100% likelihood of engine failure, emphasizing the critical role of torque margin in predicting engine health. Additionally, the parameters OAT, PA, and torque margin exhibit distinct distribution patterns between healthy and faulty engines. These differences in distribution further emphasize the significance of these parameters in distinguishing engine health states, providing valuable insights into the operational conditions that correlate with engine failures.

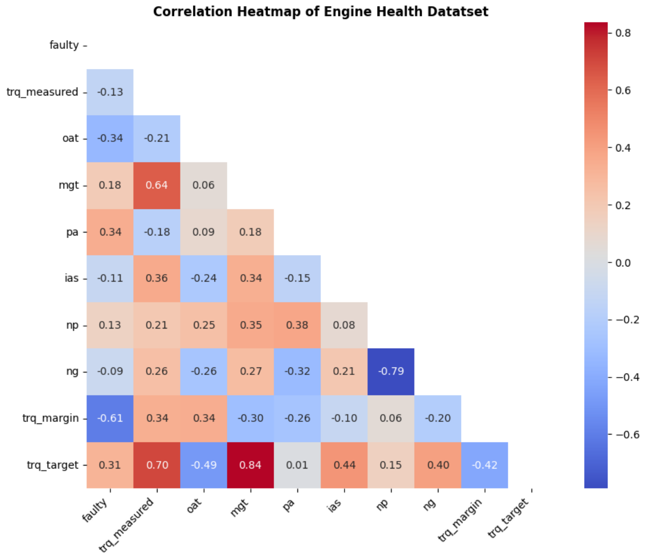

Figure 2 Correlation heatmap of engine health condition training dataset

Analyzing the correlation heatmap of the engine health dataset, as shown in Figure 2, reveals significant insights into the parameters that influence the binary 'Faulty' condition. Notably, 'Faulty' shows a moderate negative correlation with outside air temperature at -0.34, suggesting that higher temperatures might be linked with fewer faults, or conversely, cooler temperatures are associated with increased faults. Furthermore, there is a strong negative correlation of -0.61 with torque margin, suggesting that lower torque margins are significantly associated with engine faults. It can be observed that neither torque measured, nor torque target show significant correlation with engine health condition directly. However, the toque margin serves as a critical indicator, underscoring that while direct measures of torque may not reflect engine health, the torque margin plays a vital role in predicting engine health condition. This distinction emphasizes the importance of monitoring torque margin as a more sensitive and telling metric for assessing engine condition and potential failures. Additionally, the correlation between 'Faulty' and PA (Available Power) is 0.34, indicating a positive relationship where higher available power is associated with an increased likelihood of engine filature.

While these primary parameters show significant impacts, other variables such as IAS (Indicated Airspeed) and NP (Net Power) also exhibit correlations that, though minor, suggest a complex interaction affecting the engine's healthy condition. Multi-collinearity was observed in several parameters, indicating that the relationship between these variables and engine health could be nonlinear and interdependent. These interactions may complicate the predictive modeling but also offer deeper insights for more robust fault prediction and prevention strategies. These insights underscore the importance of closely monitoring temperature, torque settings, and available power to predict and prevent potential faults in engine operations. This comprehensive analysis also suggests the necessity of considering a wider range of operational parameters to fully understand and optimize engine health and performance.

3. METHODOLOGY

To classify the health condition of helicopter turboshaft engines, we initiate the process with Exploratory Data Analysis (EDA) to thoroughly examine and understand each feature of the dataset. Engine health is evaluated using the torque margin, defined as the percentage difference between the measured and target torque values. We employ Gaussian Process Regression (GPR) to predict the target (design) torque based on the turboshaft engine data. To efficiently train the GPR model, we introduce a space-filling strategy through sequential sampling, which conditionally selects a representative subset of the data, ensuring comprehensive coverage of the input space without using excessive training data. The GPR model not only predicts the target torque but also provides confidence intervals for these predictions. Using the predicted target torque, we calculate the torque margin, which serves as a key indicator of engine performance and health. This torque margin is then added as a new feature to the dataset. To classify the engine’s condition as either faulty or not, we develop an ensemble learning method. This involves creating four ML models to predict the binary class, followed by the implementation of a stacking architecture that trains on the residuals to improve prediction accuracy. This structured workflow, illustrated in Figure 3, ensures a robust and reliable machine-learning framework for helicopter turboshaft engine health monitoring.

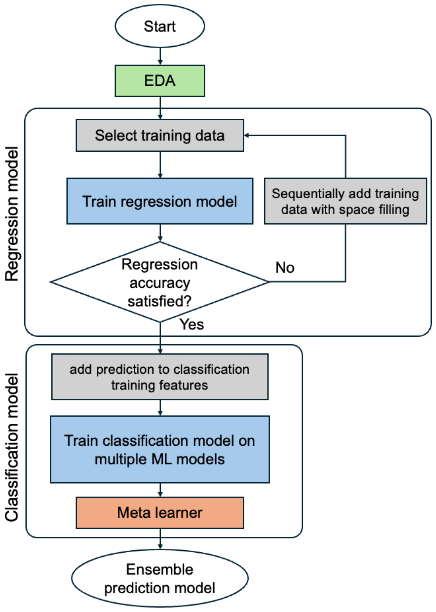
Start
EDA
Regression model
Select training data
Train regression model
Sequentially add training data with space filling
Regression accuracy satisfied?
No
Yes
Classification model
add prediction to classification training features
Train classification model on multiple ML models
Meta learner
Ensemble prediction model

Figure 3 Flowchart of framework development

3.1. Regression with Space Filling

In our methodology, we focus on selecting an appropriate probability density function (PDF) for our predictions. GPR provides predictive mean and variance estimates, making it ideal for uncertainty quantification. Since GPR inherently treats each prediction as a distribution rather than a single point, the normal distribution is a natural fit. Therefore, we selected the normal distribution as our PDF for predictions, leveraging the inherent probabilistic nature of Gaussian Processes. However, GPR model is typically well-suited for small datasets and face significant computational challenges when dealing with large volumes of high-dimensional data. In our case, we need to model a complex system with over 720,000 data points in a 7-dimensional space, making it impractical to train a GPR model using the entire dataset. To efficiently select a representative subset of data points, we adopted a space-filling strategy, which ensures comprehensive coverage of the input space without risking overfitting and excessive computation load. Specifically, we used the Max-min approach (Jin et al., 2002, Fillmore et al., 2022), which adaptively determines new training points by maximizing the minimum distance between a new point and all existing points. Formally, the new training point is identified as, where represents all current training points and is the L2-norm of a vector. This method ensures that the new points are located at most “blanked” area in the training space, thus uniformly sampling the design space and avoiding clustering of points. The Max-min approach is particularly advantageous when iterative model evaluation is time-consuming or infeasible, as it does not rely on model performance but purely on spatial relationships among points.

3.2. Ensemble Model for Classification

In this study, we utilize an ensemble learning approach to address a binary classification problem. Ensemble models are well-known for enhancing prediction accuracy and robustness by leveraging the strengths of multiple algorithms with diverse mechanisms. By integrating these algorithms, the model compensates for individual weaknesses and produces more reliable and generalized predictions. This method allows for better handling of varying data patterns and reduces the risk of overfitting, ultimately improving overall performance. The main advantage of ensemble methods lies in their ability to reduce variance, bias, or both, depending on the specific algorithms used.

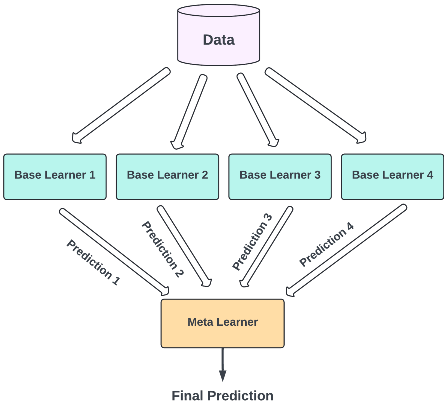

Figure 4 Flowchart of ensemble learning model

As shown in Figure 4, the ensemble model designed for this study consists of a set of base learners and a meta learner, strategically selected to optimize classification performance. The base learners include Convolutional Neural Networks (CNN), Multi-Layer Perceptrons (MLP), XGBoost, and AdaBoost, each contributing unique capabilities to the overall model. CNNs, known for their ability to capture spatial hierarchies, are used to process structured and image-like data. MLPs are leveraged for their universal approximation ability in handling non-linear relationships. XGBoost is included for its powerful gradient-boosting framework, offering efficiency and accuracy in structured data tasks, while AdaBoost provides adaptive boosting to improve model performance on difficult samples. At the top level, a logistic regression model serves as the meta learner, aggregating the predictions from the base models. This choice of meta learner is motivated by its simplicity and interpretability, making it an effective tool for combining outputs and generating final predictions. The ensemble architecture, thus, harnesses the complementary strengths of each learner while maintaining a balance between complexity and interpretability.

4. RESULT AND DISCUSSION

4.1. Prediction of Torque Margin

For regression, we randomly selected 2,000 points as initial training data, followed by the sequential selection of an additional 18,000 points using the Max-min criterion. The torque margin, calculated using Equation 1, is a key feature for indicating the health state of a turboshaft engine. During the training phase, we use the intermediate torque target as the output for the regression model. This choice is motivated by the direct relationship between the intermediate torque target and the input features. By predicting the target torque first and then calculating the torque margin using the equation, we can reduce information loss and better capture the relationship between input and output features, thereby reducing error. For the GPR model, we employed a Matérn kernel with a smoothness parameter $\mu = 3/2$. The Matérn kernel which is a type of covariance function particularly popular due to its flexibility in modeling data with varying smoothness. The result shown in Figure 5 demonstrate the efficacy of our approach. By employing the Max-min space-filling strategy, we ensure that our GPR model is trained on a well-distributed subset of data points, leading to more accurate predictions of the target torque. Consequently, when calculating the torque margin, the reduced error and improved capture of input-output relationships validate the effectiveness of our methodology. The prediction shows high accuracy and extremely low bias. This comprehensive approach not only enhances the predictive performance of the GPR model but also ensures robust estimation of the torque margin, which is crucial for monitoring the health state of the turboshaft engine.

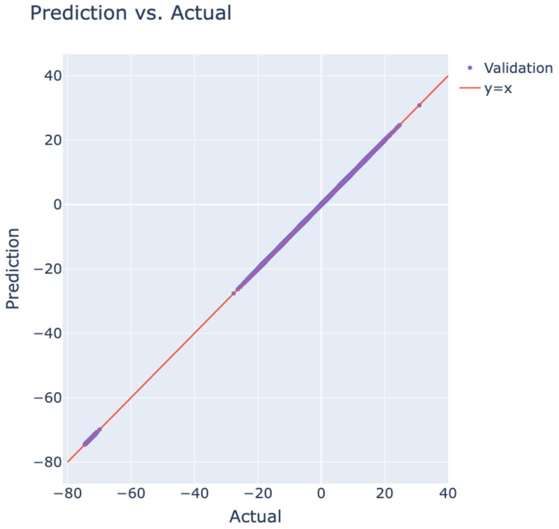

Figure 5 Regression model performance

4.2. Prediction of Engine Health Condition

In this study, 24,160 data points (3.25%) were selected as the test dataset and excluded from model training. These points were chosen because, in the subsequent task, the developed model will be used to predict the health conditions of 20 thousand data points with unknown states. To closely replicate this application scenario, we selected the same number of test points that were closest in Euclidean distance to the unknown points. If the trained model performs well on the test set, it provides strong justification for expecting stable predictions on the future unknown data points. However, some limitations need to be considered regarding this assumption. Although the testing dataset is selected by identifying the most relevant data points in relation to the unknown validation dataset, if the validation dataset represents a different high-dimensional operational space with distinct embedded physics, the trained model may still struggle to extrapolate the most accurate relationships.

OAT, MGT, PA, IAS, NP, NG and the torque margin obtained by previous regression process were selected as model input and the final output is the binary class of engine health condition. For both the MLP and 1D CNN models, the Adam optimizer is employed with a learning rate of 0.001 and a loss function of binary cross-entropy. The MLP model features three hidden layers with 50, 100, and 50 neurons, using the ReLU activation function. In contrast, the 1D CNN architecture consists of two convolutional layers with 32 and 64 filters, respectively, each using a kernel size of 3, followed by a flattening layer and a dense layer with 50 neurons. The output layers for both the CNN and MLP model have a single neuron with a sigmoid activation function for binary classification. Early stopping is implemented for both models with a patience of 10 to automatically determine the best number of training epochs, allowing both models to effectively learn patterns in the data. In the XGBoost model, a learning rate of 0.1, 500 estimators, and a maximum tree depth of 5 are configured, with the evaluation metric set to 'logloss.' For the AdaBoost model, a decision tree base estimator with a maximum depth of 4 is used, along with 300 estimators and a learning rate of 0.1, applying the 'SAMME' algorithm for ensemble learning. The logistic regression model is configured with a regularization strength parameter CCC set to 0.9, using the 'liblinear' solver for optimization, and allowing a maximum of 400 iterations for convergence. The random state is fixed at 42 to ensure reproducibility of results.

During the model training and testing process, we observed a noticeable impact of the total number of training data points on model performance. The subset of the training dataset was also selected based on the rule of filtering points with the smallest Euclidean distance at certain level. Figure 6 to Figure 8 illustrate the model's performance scores, number of wrong prediction and number of false negative prediction on the test dataset across different input sizes used in several experiments, clearly highlighting how varying the dataset size influenced predictive accuracy. All three plots demonstrate that when the training dataset size reaches approximately 190,000, the trained model achieves the highest prediction accuracy, with the lowest prediction error and the fewest false negatives. This condition may indicate another data imbalance concerning the operational space, suggesting that a large portion of the training dataset may consist of data points from a different operational space than the real unknown data. Training the model on too many biased data points could lead to overfitting, reducing the model's generalization ability. Therefore, it is crucial to select the most appropriate portion of the original full dataset as the realistic training dataset.

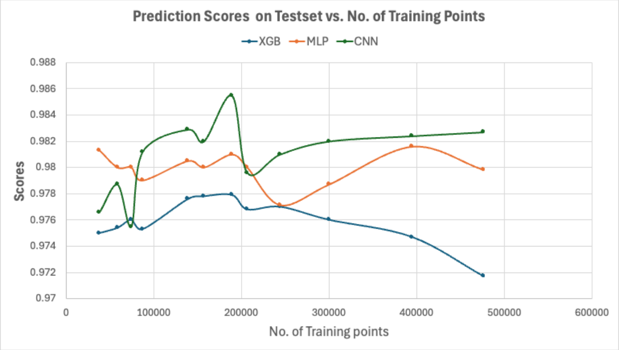

Figure 6 Prediction score Vs. No. of training points

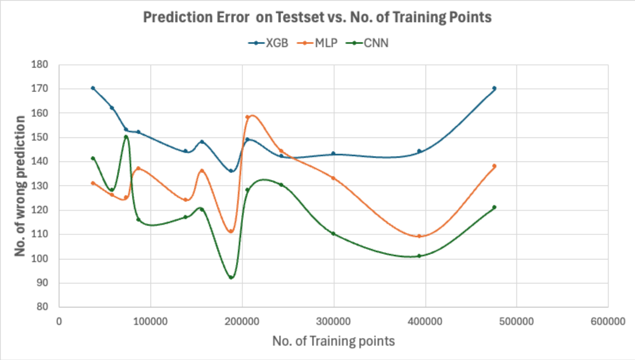

Figure 7 Prediction errors Vs. No. of training points

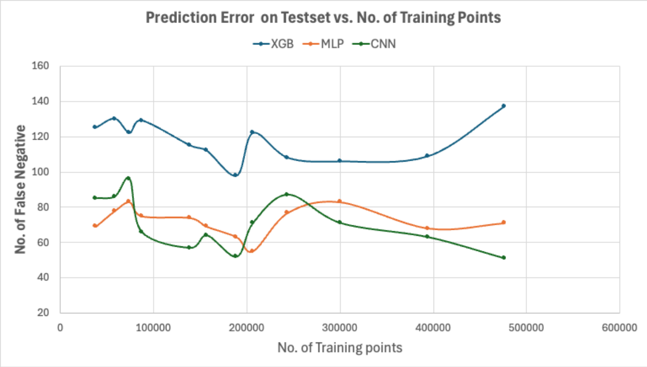

Figure 8 No. of false negative Vs. No. of training points

Finally, the ensemble model was trained on 189,782 filtered data points and tested on 24160 data points, as depicted in the prediction confusion matrix in Figure 9. The model achieved a high prediction accuracy of 99.52%, with a recall of 98.79%, precision of 98.71%, and an F1 score of 98.76%. These results indicate a robust predictive capability, as evidenced by the low number of misclassifications: only 56 false negatives and 58 false positives. This performance demonstrates the model's effectiveness in correctly identifying both healthy and faulty conditions, making it a valuable tool for proactive maintenance and fault detection in operational settings. However, there are 58 prediction results that reported false negatives. Since type II errors can be fatal for a helicopter engine, there is still room to improve the model to minimize these errors.

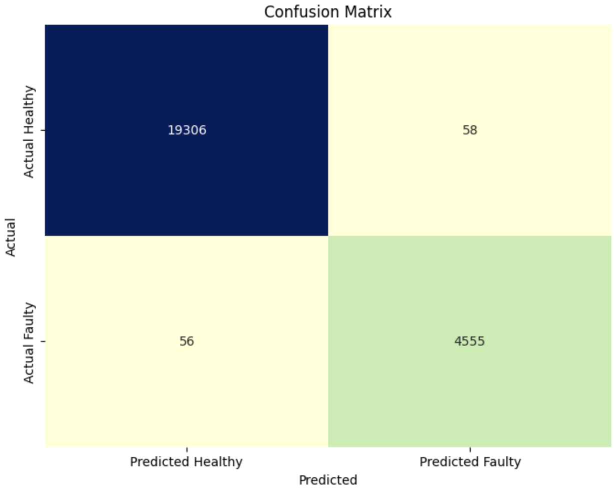

Figure 9 Confusion matrix of prediction

The Receiver Operating Characteristic (ROC) curve displayed shows an Area Under the Curve (AUC) of around 1.0, which represents an ideal scenario where the model perfectly discriminates between the positive and negative classes without any misclassification.

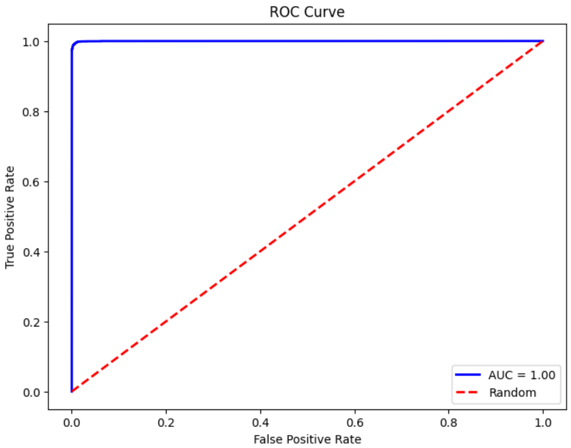

Figure 10 Receiver operating characteristic (ROC) curve

This section confirms the effectiveness of the ensemble model in predicting engine health conditions, showing robust performance across various training data sizes. The optimal results, reflected in confusion metrics and an ideal ROC curve, demonstrate the model's precision in turboshaft engine health conduction detection, making it a valuable tool for predictive maintenance in aviation turbine operations.

5. CONCLUSION

In this study, an ensemble machine learning framework is presented which delivers robust predictions of helicopter turboshaft engine health status. By combining Gaussian process regression for torque margin estimation and a stacked ensemble of classification models, the proposed approach achieves high accuracy in distinguishing between nominal and faulty engine conditions. The results demonstrate the superior generalization and reliability of the ensemble method, providing precise torque margin estimates to enable informed predictive maintenance decisions. The use of a space-filling strategy for efficient data sampling ensures the scalability of the framework to large, high-dimensional datasets. Future work should explore the generalization of the model across a broader range of engine types and the integration with real-time monitoring tools to enhance the practical utility of this approach.

REFERENCES

Elasha, F., Li, X., Mba, D., Ogundare, A., & Ojolo, S. (2021). A novel condition indicator for bearing fault detection within helicopter transmission. *Journal of Vibration Engineering & Technologies*, *9*, 215-224.

Achouch, M., Dimitrova, M., Ziane, K., Sattarpanah Karganroudi, S., Dhouib, R., Ibrahim, H., & Adda, M. (2022). On predictive maintenance in industry 4.0: Overview, models, and challenges. *Applied Sciences*, *12*(16), 8081.

Wu, Z., Zeng, J., Hu, Z., & Todd, M. D. (2023). Optimization of unmanned aerial vehicle inspection strategy for infrastructure based on model-enabled diagnostics and prognostics. *Mechanical Systems and Signal Processing*, *204*, 110841.

Daouayry, N., Maisonneuve, P. L., Mechouche, A., Scuturici, V. M., & Petit, J. M. (2018). Predictive maintenance for helicopter from usage data: application to main gear box.

Schwabacher, M. (2005). A survey of data-driven prognostics. *Infotech@ Aerospace*, 7002.

Wang, T., Yu, J., Siegel, D., & Lee, J. (2008, October). A similarity-based prognostics approach for remaining useful life estimation of engineered systems. *In 2008 international conference on prognostics and health management* (pp. 1-6). IEEE.

Wu, Z., Fillmore, T. B., Vega, M. A., Hu, Z., & Todd, M. D. (2022). Diagnostics and prognostics of multi-mode failure scenarios in miter gates using multiple data sources and a dynamic Bayesian network. *Structural and Multidisciplinary Optimization*, *65*(9), 270.

Coble, J. B., & Hines, J. W. (2008, October). Prognostic algorithm categorization with PHM challenge application. In *2008 International Conference on Prognostics and Health Management* (pp. 1-11). IEEE.

Surucu, O., Gadsden, S. A., & Yawney, J. (2023). Condition monitoring using machine learning: A review of theory, applications, and recent advances. *Expert Systems with Applications*, *221*, 119738.

Nelson, W., & Culp, C. (2022). Machine learning methods for automated fault detection and diagnostics in building systems—A review. *Energies*, *15*(15), 5534.

Wang, J., Jing, H., Ozbayoglu, E., Baldino, S., Zheng, D., & Li, X. (2024, June). Enhancing Well Kick Classification in Drilling Operations Using a Novel PCA-Based Machine Learning Approach. In *ARMA US Rock Mechanics/GeomechanicsSymposium* (p. D031S034R001). ARMA.

Zheng, D., Wang, J., Jing, H., Ozbayoglu, E., Silvio, B., & Jakaria, M. (2024, June). Identifying the Robust Machine Learning Models to Cement Sheath Fatigue Failure Prediction. In *ARMA US Rock Mechanics/Geomechanics Symposium* (p. D042S058R006). ARMA.

Mian, Z., Deng, X., Dong, X., Tian, Y., Cao, T., Chen, K., & Al Jaber, T. (2024). A literature review of fault diagnosis based on ensemble learning. *Engineering Applications of Artificial Intelligence*, *127*, 107357.

Jin, R., Chen, W., & Sudjianto, A. (2002, January). On sequential sampling for global metamodeling in engineering design. In *International design engineering technical conferences and computers and information in engineering conference* (Vol. 36223, pp. 539-548).

Fillmore, T. B., Wu, Z., Vega, M. A., Hu, Z., & Todd, M. D. (2022). A surrogate model to accelerate non-intrusive global–local simulations of cracked steel structures. *Structural and Multidisciplinary Optimization*, 65(7), 208.