Identifying Key Factors in Turbofan Engine Health Degradation using Functional Analysis

Declan P. Mallamo[1], Michael H. Azarian[2], Michael G. Pecht[3]

[1,2,3]*University of Maryland, College Park, MD 20742 USA*

dmallamo@umd.edu
mazarian@umd.edu
pecht@umd.edu

ABSTRACT

This study introduces the Elastic Sparse Functional k-Nearest Neighbors approach, a predictive health monitoring framework specifically tailored for turbofan engines. This method begins by transforming time-series data into a standardized universal flight domain, which is further optimized through elastic registration for alignment across varying flight regimes. Standard scaling is employed as a preprocessing step, setting the stage for feature dimensionality reduction via Functional Principal Components Analysis. To pinpoint the features that most significantly impact engine health, the method leverages Orthogonal Matching Pursuit in conjunction with k-Nearest Neighbors to build a sparse regression model. The model's performance is assessed using root mean square error on test cases derived from the NCMAPSS DS02 dataset. Recommendations are given based on the interpretive results relating to targeting data collection and formulating hypotheses for root cause analysis.

1. INTRODUCTION

Today, predictive health monitoring is needed to ensure the operational integrity and safety of turbofan engines in the aerospace industry. Effective predictive modeling can help predict rare faults and failures that have occurred in the fleet, which has the potential to prevent future failures, reduce the time necessary for failure identification, reducing 'no-fault-found' incidents, improving troubleshooting for new problems, reducing unscheduled maintenance interruptions, and shifting to a more cost efficient conditioned-based maintenance strategy to help to forecast engine degradation which minimizes unscheduled maintenance and extends the service life of the engines.

The challenge of training efficient fault prognostics models in any medium, in accessing representative and informative

data. Poor data quality such as incompleteness, inconsistency or inaccuracies and large data volume can hinder effective analysis. Bias can be introduced through data collection and labeling, leading to difficulty in generalizing results to other systems due to large data variation.

There have been surveys performed on the benefits and challenges in implementing prognostic approaches (Sun, Zeng, Kang & Pecht, 2012). An informative review was written on physics-based and data-driven methods (An, Kim & Choi, 2015). For failures associated with most mechanical systems, there are physical or chemical processes associated with the failure mechanisms, such as corrosion, overstress or wearout. In cases where sensory information is limited, physical models can help fill in the gaps in the analysis. But in cases where physical models must be inferred, and sensor information is limited, data-driven models can still be developed to make remaining useful life (RUL) predictions using a state of health from indirect sensor measurements. State-of-health indications have been derived from vibration (Wang, Miao & Kang, 2009) (Lee, Azarian & Pecht, 2020), temperature (Lall & Thomas, 2017), and oil debris mass spectroscopy signals (Dupuis, 2010), and using this framework, the underlying degradation process of a physical system may be modeled. A derived mathematical index is calculated using a relationship between difference of the systems parameters from nominal conditions to the point of failure.

Li, Li, Zuo, Zhu and Shen (2022) found that employing domain adaptation to train a model on a well-labeled source dataset and adapting it for an unlabeled target dataset, data is aligned at both feature and semantic levels to improve generalization and accuracy. The dual-level alignment process may introduce complexity that could hinder straightforward interpretability of the model's predictions (Li et al., 2022). Biggio, Weiland, Chao, Kastanis & Fink (2021) presented Deep Gaussian process like Deep Sigma Point Processes (DSPPs) to enhance standard Gaussian Process models by combining deep learning architectures with probabilistic methods to provide both accurate predictions

and uncertainty quantifications (Biggio et al., 2021). There has been research focused on building degradation-based prognostics for turbofan engines using a state space model (SSM) to describe the system level latent degradation dynamics when only various performance data are available (Sun, Zuo, Wang and Pecht, 2012) (Saxena, Goebel, Simon & Eklund, 2008).

The present study introduces the Elastic Sparse Functional k-Nearest Neighbor (ESF-kNN) method, a specialized pipeline for predictive health monitoring in turbofan engines. Initially, time-series data is transformed into a standardized universal flight domain, paving the way for aligning various operational regimes through elastic registration. This alignment employs a time-warping function to ensure coherent phase relationships across different flights. Following alignment, standard scaling is applied to normalize the data, facilitating the application of Functional Principal Components Analysis (FPCA) for dimensionality reduction. The FPCA yields principal components that encapsulate the primary modes of variability across the flights. After FPCA, Orthogonal Matching Pursuit (OMP) is utilized to select a sparse set of these principal components, each corresponding to distinct engine health parameters and specific aligned flight regimes. OMP operates by iteratively choosing features orthogonal to the model's residuals, thereby each contributing unique, non-redundant information. This sparse set of features is then used to construct a k-nearest neighbor (k-NN) regression model, which employs discrete 3rd-order B-spline interpolations at consistent points along the universal flight domain as features. The model offers a balance between predictive accuracy and computational efficiency. By narrowing the focus to these sparse and critical features, ESF-kNN not only enhances prediction quality but also affords crucial insights into the variables and flight regimes most indicative of engine health. This makes the method particularly adept at handling test scenarios that feature operational regimes not well-represented in the training set.

1.1. Contributions

This study introduces the ESF-kNN approach, aiming to improve predictive health monitoring in turbofan engines. One of the key contributions is the effective use of temporal domain normalization and elastic functional data registration for preprocessing. These steps are designed to tackle the specific issue of misaligned time-series data, a common challenge in turbofan engine monitoring. Utilizing FPCA for dimensionality reduction, the method addresses the high-dimensionality seen when adapting OMP to work effectively with time-series data. This is particularly relevant as OMP is generally not well-suited for non-aligned time-series data. Combined with k-NN, the ESF-kNN framework aims to provide a balanced trade-off between predictive accuracy and computational efficiency. The method is validated using the NC-MAPSS DS02 data set, showing that it can be used to handle unobserved flight regimes.

In Section 2, a discussion is given for the rationale for converting the time domain to a standardized flight interval, as well as the advantages of elastic registration for aligning functional data. An examination of the role of standard scaling in data preprocessing and the application of FPCA for reducing feature dimensions is then given followed by the combined use of OMP and k-NN for effective feature selection. The final segments focus on validating the methodology using Root Mean Square Error metrics on select test cases from the NCMAPSS DS02 dataset and conclude with discussions on the model's generalizability and interpretability.

2. DATA AGGREGATION AND PROGNOSTICS

This work develops a prognostic method using data from NASA's Prognostic Center of Excellence. The data was generated using a C-MAPSS model simulating a turbofan engine's lifecycle with high pressure turbine degradation, using real flight data as inputs (Chao, Kulkarni, Goebel & Fink, 2021). The DS02 dataset has limits the failure modes to only high-pressure turbine efficiency and low-pressure turbine efficiency and flow modifications. Fig. 1 shows a picture of a large turbofan engine, highlighting the recorded input and output parameters.
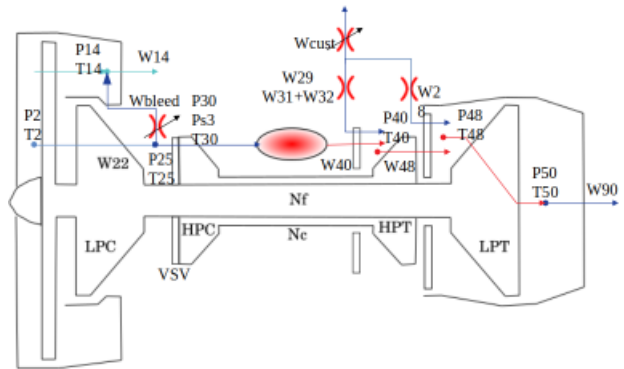

Fig. 1. Turbofan engine graphic showing positions and directions of flow, temperature and pressure measurements used in the analysis.

2.1. NC-MAPSS dataset

This work develops a prognostic method using data from NASA's Prognostic Center of Excellence. The data was generated using a C-MAPSS engine model simulating the turbofan engine's lifecycle with high pressure turbine degradation (Chao et al., 2021). Real flight conditions were recorded on a commercial jet using NASA's DASHlink system and fed as inputs to the NC-MAPSS model (Frederick, DeCastro & Litt, 2007). The simulation covers a flight envelope, concerning altitude (Alt), Mach number (M), throttle-resolver angle (TRA), and total fan inlet temperature (T2). A complete set of output parameters with units can be found in Table 1.

Table 1. Output parameters and units.

Nomenclature	Symbol	Units
HPT efficiency modifier	HPT_eff_mod	%
LPT efficiency modifier	LPT_eff_mod	%
LPT flow modifier	LPT_flow_mod	%
Fuel flow	W_f	pph
Total temperature at fan inlet	T_2	°R
Total temperature at LPC outlet	T_{24}	°R
Total temperature at HPC outlet	T_{30}	°R
Total temperature at LPT outlet	T_{50}	°R
Pressure at fan inlet	P_2	psia
Total pressure in bypass duct	P_{15}	psia
Total pressure at HPC outlet	P_{30}	psia
Physical fan speed	N_f	rpm
Physical core speed	N_c	rpm
Engine pressure ratio (P50/P2)	epr	---
Static pressure at HPC outlet	P_{s30}	psia
Fuel equivalency ratio (W_f/P_s30)	φ	pps/psi
Corrected fan speed	NR_f	rpm
Corrected core speed	NR_c	rpm
Bypass ratio	bpr	---
Fuel equivalency ratio for the burner	φ_B	---
Bleed enthalpy	ht_{Bleed}	---
Demanded fan speed	$N_{f,dmd}$	rpm
Corrected demanded fan speed	$N_{Rf,dmd}$	rpm
HPT cooling bleed	W_{31}	lbm/s
LPT cooling bleed	W_{32}	lbm/s
Total temperature at HPT outlet	T_{48}	°R
Fan stall margin	SmFan	---
LPC stall margin	SmLPC	---
HPC stall margin	SmHPC	---

The naming conventions and parameters in turbofan engines serve specific roles in monitoring and diagnosing the system's overall health and efficiency. Parameters prefixed with 'T' are temperature measurements at various engine sections. For example, T_2 is the total temperature at the fan inlet, crucial for air intake conditions, while T_{40} and T_{48} provide temperatures at the High-Pressure Compressor (HPC) and High-Pressure Turbine (HPT), vital for thermal efficiency. Similarly, parameters with a 'P' prefix indicate pressure levels. P_2 measures the pressure at the fan inlet, essential for air intake efficiency. P_{25} and P_{50} offer insights into the effectiveness of the Low-Pressure Compressor (LPC) and engine exit pressure, respectively. Airflow and fuel flow rates are denoted by the 'W' prefix. W_{22} and W_{40} are the airflow rates at the LPC and HPC, necessary for optimal combustion. W_{48} signifies the airflow at the HPT, which is critical for turbine performance. W_{90} typically captures bleed flows for bypass or customer use. Lastly, rotational speeds are indicated by 'N', where N_f and N_c represent the physical fan and core speeds, respectively. These speeds are indicators of the engine's rotational performance, affecting compression, combustion, and overall system efficiency.

2.2. Transformation from Time Domain to Universal Flight Domain

The transformation of time series data into a universal flight domain, spanning from 0 to 100, serves as a normalization step in the preprocessing pipeline for turbofan engine health prognosis. This temporal standardization allows for a harmonized analytical framework across flights with differing durations and operating regimes. To facilitate this transformation, each flight's time series is resampled to conform to this universal temporal scale. In this study, B-spline interpolation is employed to carry out this resampling process. The B-spline interpolation has the advantage of being piecewise-defined, meaning that each polynomial piece is defined over a subinterval of the total interval of t. A definition of B-splines can be found below in Eq. (1).

$$f(t) = \sum_{i=0}^{n} P_i B_{i,k}(t)$$
$$B_{i,1}(t) = \begin{cases} 1, & if, t_i \leq t \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases}$$
$$B_{i,k}(t) = \frac{t-t_i}{t_{i+k-1}-t_i} B_{i,k-1}(t) + \frac{t_{i+k}-t}{t_{i+k}-t_{i+1}} B_{i+1,k-1}(t) \quad (1)$$

B-splines are particularly well-suited for this task due to their ability to approximate complex data patterns with a controllable degree of smoothness, thereby minimizing the risk of overfitting or underfitting the data. This high-fidelity representation ensures that the inherent variability and trends in the original time series are also captured in the resampled functional data.

2.3. Using Elastic Registration on Interpolated Functional Curves for Enhanced Predictive Modeling

Elastic registration identifies a warping function that, when applied to the time domain, aligns curves in such a way as to minimize a loss function (Srivastava & Klassen, 2016). It is especially useful in cases where the 'when' is less important than the 'what'. Suppose f(t) and g(t) are two functions to be registered, the goal of elastic registration is to identify a warping function, γ(t), that minimizes a distance metric, D, between f(t) and g(γ(t)), identified below in Eq. (2).

$$\gamma^*(t) = argmin_\gamma D\left(f(t), g(\gamma(t))\right) \qquad (2)$$

Where the boundary conditions, $\gamma(0)=0$ and $\gamma(1)=1$, ensure that the endpoints are aligned, and $\gamma(t)$, is a monotonically increasing function. The warping function is applied such that $\gamma'(t)>1$ increases the time increments and $\gamma'(t)<1$ decreases them. In this research, the Square Root Velocity (SRV) function is differentiable, thus allowing for shape comparison using the L^2-norm and can be solved using dynamic programming. Using the SRV warping function, $q(t)=\sqrt{f'(t)}$, the optimization problem found in Eq. (3) becomes:

$$Elastic\ Distance^2 = argmin_\gamma \int_0^1 (||q(t)-(q_g \circ \gamma)(t))||)^2 dt \quad (3)$$

Where $(q_g \circ \gamma)(t)$ signifies the composition of the SRV function evaluated for $g(t)$ and the warping function. Once the warping function is applied to the time domain of your interpolated data, the space is re-gridded to form a new set of discrete points.

2.4. Generalizing State of Health Labels using Geometric Mean.

To combine the different state of health (SoH) labels to create a prognostic to work with three types of faults seen in the N-CMAPSS DS02 dataset, LPT and HPT efficiency modifications and LPT Flow modification, the geometric mean was used to ensure the universal SoH metric is high when all individual SoH metrics are high low when either one is low. This is seen in Eq. (4) below.

$$Universal\ SoH = (\textstyle\prod_{i=1}^{n} SoH_i)^{\frac{1}{n}} \quad (4)$$

2.5. Standard Scaling of Multivariate Cross-Sectional Data.

The scales of the different parameters associated with multivariate cross-sectional data can disproportionately influence machine learning algorithms, and standard scaling ensures all the variables contribute equally to the resulting estimate. The process of standard scaling involves subtracting the mean and dividing by the standard deviation for each column in the dataset as seen in Eq. (5).

$$z = \frac{x-\mu}{\sigma} \quad (5)$$

Subtracting the mean and dividing the result by the standard deviation causes the data to have unit spread over the distribution.

2.6. Extending the Features using Derivatives

The forward finite difference method, seen in Eq. (6) below, was used to approximate the derivatives of each of the gas path parameter functionals to extend the initial features and increase the available information.

$$f'(x) = \frac{f(x+h)-f(x)}{h} \quad (6)$$

2.7. Dimensionality Reduction with Functional Principal Component Analysis

Functional Principal Component Analysis (FPCA) serves as a multivariate analytical technique that models variables over time through a set of functions, thereby illuminating underlying functional relationships. Unlike traditional multivariate Principal Component Analysis (PCA), which deals with observational data, FPCA works with functional data to extract features as functional principal components. These components are orthogonal linear combinations of continuous functions that capture significant variations in the data.

The mathematical foundation of FPCA lies in the concept of inner products. For Euclidean space, the inner product $<x,y>$ $= x^T y$ and for L^2-space $<x,y>=\int x(t)y(t)dt$. Using this form of the inner product allows for the generalization of the multivariate principal component problem in Eq. (7), below.

$$argmin_{\xi_1} \textstyle\sum ||x_j - <x_j,\xi_1>\xi_1|| \quad (7)$$

For multivariate data, the inner product is a piece-wise linear combination of features. Finding the first principal component amounts to solving:

$$max_{\beta^T\beta=1} \frac{1}{N}\beta^T X^T X\beta \quad (8)$$

The solutions are found by solving the eigenvalue problem, $V\beta = \lambda_1 \beta$, where $V = \frac{1}{N}(X^T X)$ is the sample covariance matrix.

In the functional setting, the covariance operator V_f replaces the covariance matrix, defined as:

$$V_f(s) = \int_T v(s,t)f(t)d(t) \quad (9)$$

Where $v(s,t) = \sum_{i=1}^{N} x_i(s)x_i(t)d(t)$. The maximization problem takes on the following form:

$$\int_T v(s,t)\beta_j(t)d(t) = \lambda_j\beta_j(s), j = 1,2, \dots \quad (10)$$

This equation can be solved using a spectral decomposition method seen in Eq. (11),

$$V_f = \textstyle\sum_{i=1}^{\infty}\lambda_i(s) < f, \beta_j > \beta_j \quad (11)$$

Here, $\{\beta_j\}$ for an orthonormal basis, each with their respective eigenvalue, λ_i. By extending the concept of multivariate principal component analysis, where a high dimension can be represented as a low dimension set of principal components to functionals, an infinite number of dimensions can be reduced to a small finite set which is a significant improvement (Hong, 2020).

2.8. Integrated k-NN-OMP for Turbofan Engine State-of-Health Monitoring

Once the FPCA is used to identify the principal components accounting for data variability, OMP is then employed to find a sparse set of these principal components and corresponding universal flight regimes that have the most impact on engine

performance and degradation. These identified regimes and components are further used as inputs to a k-NN regression model to predict the engine's SoH. This streamlined approach not only enhances prognostic performance but also provides insights into critical flight regimes and engine parameters, enabling the development of less intrusive and more efficient health monitoring systems.

2.8.1. k-NN Regression

K-NN regression is a non-parametric method used for predicting continuous output variables. Given a query point and a set of observed data points, k-NN identifies the k-nearest neighbors in the feature space and averages their output values for prediction. The number of neighbors, k, was found using Monte Carlo cross-validation where random training/testing Pareto partitions were used to identify the optimal value of k. The k-NN regression method can be represented using Eq. (12) below.

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i \qquad (12)$$

Where $\hat{y}$, is the predicted output made up of the averages of the k-nearest neighbors.

2.8.2. OPM algorithm

The core objective of the OPM algorithm is to address an under-determined inverse problem, where the aim is to find an unknown variable x that satisfies Ax = b, given a "short and fat" measurement matrix A and observed data b. In practice, the observed data b is often contaminated with noise, denoted by ϵ, making the problem more challenging. The inverse problem is to work backward to discover what x must have been to yield b when multiplied by A, as opposed to the generally simpler forward problem of finding b given A and x. In this under-determined setting, an infinite number of solutions for x exist. However, the focus is on finding an x with the fewest non-zero elements, which equates to solving a computationally difficult (NP-hard) optimization problem that aims to minimize the number of non-zero elements in x while still satisfying Ax = b.

2.9. Hyper parameter optimization

A method of hyperparameter optimization was used to avoid bias in the model and overfitting to the training data. The data was split into training and testing sets, and the model was trained recursively using Monte Carlo cross-validation where random partitions of the training data are to find the optimal set of hyperparameters that result in the lowest empirical error.

2.10. Remaining Useful Life Estimation

Once a state-of-health estimate is reached it can be combined with the cycle count data to fit a remaining useful life model, by fitting an exponential decay model to SoH measurement below in Eq. (13).

$$RUL = a\, e^{b\, SoH} + c \qquad (13)$$

2.11. Spearman Correlation Coefficient

The Spearman rank-order correlation coefficient, often denoted as rs, is a non-parametric measure of the strength and direction of the relationship between two variables. Unlike Pearson's correlation, which requires the data to be normally distributed and linearly related, Spearman's correlation only assumes that the data can be ranked, given by the following formula in Eq. (14).

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \qquad (14)$$

Where n is the number of ranked scores and di is the distance between incremental rankings.

2.12. Algorithm Summary

In summary, the recorded engine parameters are used to create a SoH estimate by combining multiple turbine and flow efficiency scores from nominal together using the geometric mean. Each time series data is resampled using B-spline interpolations to convert to a universal flight domain that goes from (0,100). Flight regimes for each parameter are aligned to fit with functional representations and combined with the SoH estimate to make predictions using k-NN method. An orthogonal matching pursuit algorithm determines optimal to improve performance, run-time, data management and reduce noise, allowing for easier interpretation of results by focusing on important variability trends in the parameter waveforms. The overall method of evaluating the prognostic is found below in Fig. 2.
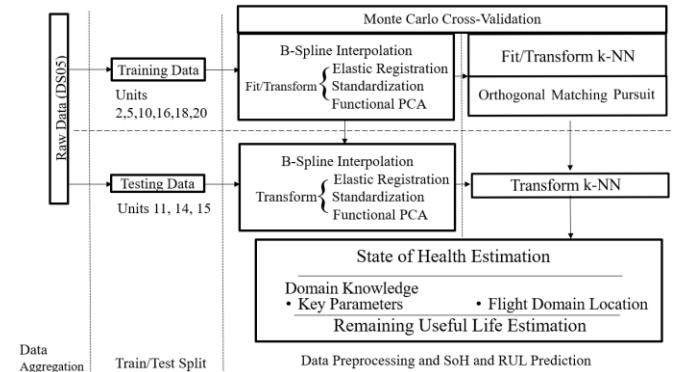

Fig. 2. Diagram showing overall prognostic method to acquire parameters of interest to target failure analysis and state of health monitoring.

3. RESULTS

The following section presents results obtained from the ESF-kNN approach applied to the NC-MAPSS DS02 dataset. The discussion begins with the impact of preprocessing steps, including universal flight domain standardization and elastic registration, on feature alignment and standardization.

Attention then shifts to the feature selection capabilities of OMP, highlighting its effectiveness on functionally aligned time-series data. Finally, the predictive accuracy of the k-NN regression model is evaluated, emphasizing its robustness and interpretability in the context of turbofan engine health prognosis.

3.1. Summary of the Key Features after Elastic Registration, Standardization, and FPCA.

In Fig. 3, a matrix displays the processed values for each parameter and its derivative, each color-coded according to their SoH labels. Vertical red dashes within the universal flight domain highlight the regions that are most influential in determining engine health, as identified by the OMP method.
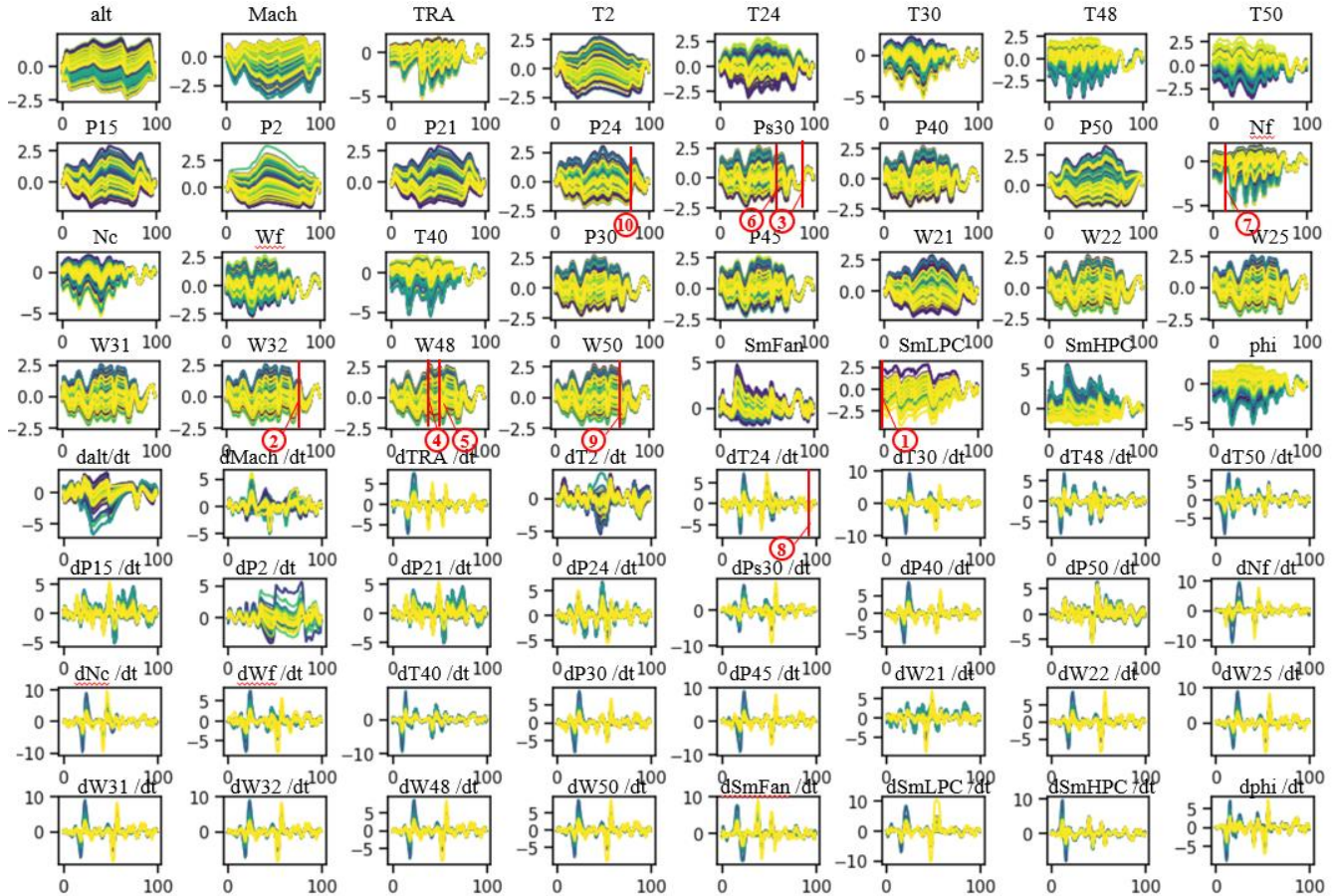

Fig. 3. Conditioned features for the testing data with red vertical marks showing the key features and rankings.

In Fig. 4, individual SoH estimates for testing unit #11, 14, and 15 are presented, each accompanied by a one-standard-deviation empirical confidence interval. Unit #11 exhibits the highest performance which could be explained by having flights that were closest in similarity to the training data in terms of altitude and flight duration.
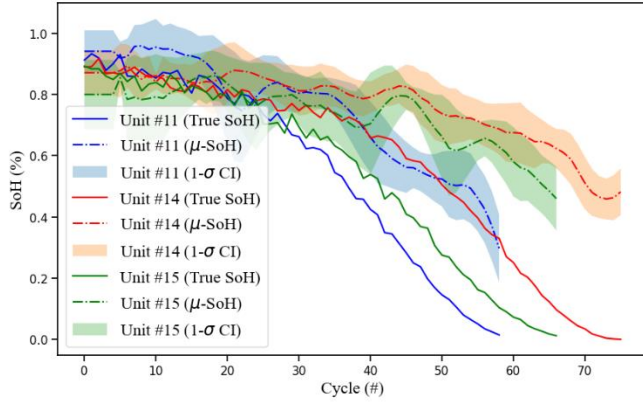

Fig. 4. State of health estimates for the different test data units in DS02 dataset. The solid lines indicate the true SoH score found using Equation 4 that combines the individual low- and high-pressure turbine efficiency and flow health scores together and normalizes them. The dashed lines signify the predictions with 1 standard deviation empirical confidence intervals.

In Table 2, the root mean squared error for the different SoH estimates can be found for both the training and testing data from the NCMAPSS DS02 dataset. The ESF-kNN method demonstrates robust performance in predictive health monitoring of turbofan engines, as evidenced by its root mean squared error (RMSE) of the RUL predictions.

Table 2. Model Performance Metrics ESF-kNN method for the training and testing data in the NCMAPSS DS02.

Unit Label	RMSE of SoH (%)
Training Data	
#2	18+/-4
#5	15+/-1
#10	15+/-3
#16	19+/-6
#18	24+/-6
#20	24+/-5
Testing Data	
#11	21+/-4
#14	22+/-2
#15	22+/-7

To contextualize the performance of the ESF-kNN method, Table 3 provides a comparative analysis of the RMSE on the NCMAPSS DS02 testing data, along with results from existing works in the literature. From looking at the performance in terms of RMSE, it is evident that the ESF-kNN method is competitive to less interpretable methods

found in the literature which underscores its effectiveness and potential for broader application in the field of predictive health monitoring for turbofan engines.
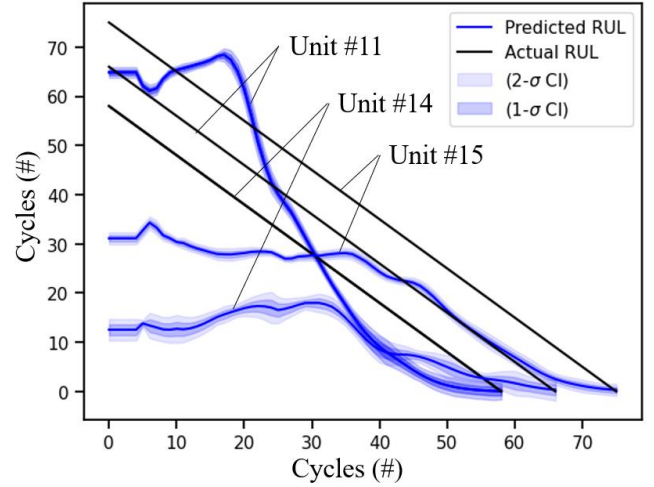

Fig. 5. Remaining Useful Life Estimates created by curving the degradation model of state of health using Eq. (13).

Table 3. RMSE Comparison of RUL estimates broken down to unit and compared to Domain-Adaptive Transformer results found in (Li et al., 2022).

Unit Label	RMSE of RUL (Cycle)
Training Data	
#2	3.26+/-0.03
#5	9.40+/-0.02
#10	7.87+/-0.06
#16	3.43+/-0.22
#18	7.12+/-0.18
#20	7.17+/-0.06
Testing Data	
#11	12.54+/-0.28
#14	21.2+/-0.45
#15	27.18+/-0.85

The method of calculating the error documented in (Chao et al., 2022), suggests that the training and testing data points were chosen such that they were interleaved throughout the flights, and not separated into to designated training and testing units, per the recommendation in (Chao et al., 2021). The original data challenge separates the original training and testing data in term of the units, which introduces difficulty associated with having many of the flights in the testing data that are non-representative in the training data, due to the different flight characteristics. For this comparison, the error associated with the application of the domain adaptive method was derived by averaging the predictions error for each of the 7 subset domains adapted onto the DS02 dataset, (4.83+5.34+12.02+29.14+31.45+33.23+14.60)/7=18.66).

Table 3. RMSE Comparison to values recorded in the literature

Algorithm	RMSE (%)	Notes
ESF-kNN	20.30 +/- 0.53	[Proposed method]
Domain-Adaptive Transformer	18.66	(Li et al., 2022)
DSPP	7.38	(Chao et al., 2021)
MCD	7.31	(Chao et al., 2021)

By incorporating the ESF-kNN method, the performance was acquired comparable to existing methods found in the state-of the art that cannot provide an explanation into the underlying features that had the most impact on the prognostic performance.

3.2. Interpretability Insights

This section delves into the key engine parameters and flight regimes identified by the ESF-kNN method. By utilizing OMP and k-NN, the approach isolates critical features for engine health, enriching our understanding of degradation patterns across various operational conditions. Fig. 6 below highlights the relative importance of these selected features.
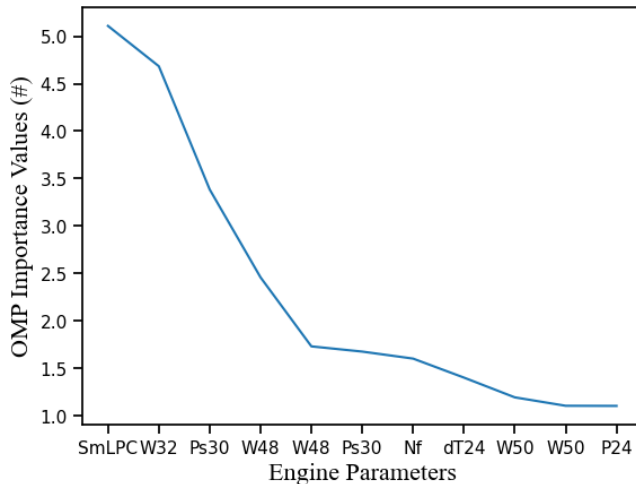

Fig. 6. A ranking of the top 9 most important features identified by OMP.

In Fig. 7 below, each of the highest scored features can be seen, centered around their respective flight domain region. The most important parameter identified was the stall margin for the low-pressure compressor, SmLPC, which was identified at the beginning of the flight. This makes sense as the Low-Pressure Compressor (LPC) is vital during the initial thrust buildup. The LPT cooling bleed, W_{32}, becomes critical at 72% into the universal flight domain, which from inspection occurs upon initial descent. This could be tied to the need for effective cooling of the engine's high-pressure components as the aircraft prepares for landing, and the operational conditions change. The static pressure at the HPC outlet, Ps30 peaks at around 86% of the universal flight domain, which is also within the descent flight mode, and shows a reiterated

importance score early on at the point of entering the descent flight phase. The HPC outlet pressure could be critical for maintaining engine stability as the plane descends and air pressure changes. The flow through the high-pressure turbine, W48 shows a high importance factor which is reiterated in the sparse feature set, reaffirming its importance in improving the prognostic performance. The features identified occur at around 40% in the universal flight domain which is in the cruise flight regime. At cruising levels, the engine's efficiency is paramount, and the flow out of the high-pressure turbine could be a key performance indicator at this point.
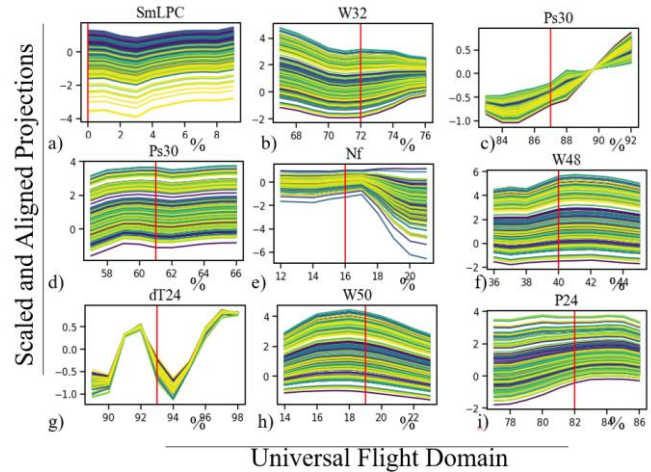

Fig. 7. Top nine key features identified through the ESF-kNN method showing relative flight domain locations. This sparsity-based method reduces the feature set from 6400 ((32 features+32 feature derivatives) × 100 interpolated points) to just nine key features identified by the vertical red lines.

The fan speed, Nf becomes prominent at 16% of the flight phase which is during the climb portion. The fan speed is critical here for generating the necessary thrust for the aircraft to maintain optimal climb angle. The low-pressure turbine output flow, W_{50} becomes significant during the climb phase right after takeoff. At this stage the engine will be at max throttle and the flow out of the low-pressure turbine could be vital for maintaining thrust and engine efficiency. The derivative of the LPC outlet temperature, dT24/dt shows a strong influence at approximately 92% of the universal flight domain. This could imply that temperature changes in the LPC outlet could be indicative of the engine's overall health, especially as it prepares for landing. Finally, the low-pressure output pressure, P_{24}, plays an important role at 82% of the universal flight domain, which is in the middle of the descent period. During the descent the engines throttles are brought back, to reduce power and prepare for landing. At this stage, managing P_{24} is crucial because it directly affects the balance between the engine's efficiency and the required thrust.

To complement the analysis correlations between the identified key features and the different state of health scores is given below in Fig. 8. The Spearman correlation matrix

provides several noteworthy insights into the relationships between different engine parameters and the general State of Health (SoH). Here, the values close to 1 or -1 indicate a strong positive or negative correlation, respectively, while values close to zero suggest weak or no correlation.
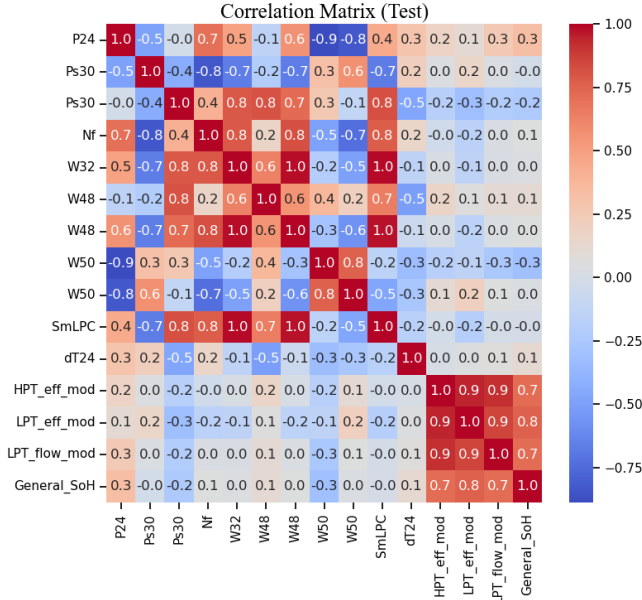

Fig. 8. Correlations between the identified key features and the different state of health scores. The rate of change of the fan inlet temperature, P_{24} and W_{50} show the greatest correlation trends with the general state of health parameter.

Upon examination of the correlation matrix for the testing data, several noteworthy observations are identified. First, the correlation between SmLPC and both W32 and W48 is almost perfect, standing at 0.99, which suggests that these two parameters are practically linearly related, and one could be used to predict the other with high accuracy. Next, the correlation between HPT_eff_mod, LPT_eff_mod, and LPT_flow_mod with General_SoH parameter is all above 0.7, which is because they were all used to create the General_SOH using Equation 4. Interestingly, P_{24} shows a negative correlation of (-0.88) with W50, and both parameters show strong correlation trends with the General_SoH parameter. Interestingly, the parameter with the largest importance score from the OMP method, the stall margin for the low-pressure compressor, taken at initial departure, SmLPC, did not show a strong direct correlation the general state-of-health score. It is important to know that while the Spearman correlation matrix offers valuable insights, it's crucial to consider that it may not fully capture the multivariate and possibly nonlinear relationships identified in more sophisticated predictive models.

3.3. Future research considerations

Advanced feature engineering could be a focal area to capture intricate patterns especially evident during non-nominal

flights. Data augmentation techniques may be developed to bridge the divergence between testing and training flights, thereby enhancing model generalizability. Temporal data trends may also be mapped through specialized time-series models to understand the incremental evolution of the state-of-health over time. Adding semantic-level information in the training step could improve overall prediction and the interpretability by reducing variability in the training and testing data. Since turbofan engines are critical safety items, more work on the integration of uncertainty quantification methods could add a layer of reliability to the predictions.

4. CONCLUSION

The developed method, termed ESF-kNN, employs a sophisticated data preprocessing pipeline that includes B-spline interpolation, elastic registration, standardization, and the computation of derivatives to transform time-series data into a functional form conducive for interpretable failure prognostics. The application of functional PCA further refines the feature set, reducing dimensionality. This processed data then feeds into a sparse k-NN algorithm, optimized using Orthogonal Matching Pursuit, to identify key parameters and flight regimes crucial for predicting Remaining RUL of turbofan engines. The study revealed key insights into the critical engine parameters affecting the overall State of Health (SoH) of turbofan engines across various flight regimes. The ESF-kNN method efficiently reduces a large 6400-parameter feature set to just nine crucial parameters, each associated with specific segments of a universal flight domain. These parameters, ranging from the stall margin in the Low-Pressure Compressor (SmLPC) to fan speed (Nf) and static pressure at the High-Pressure Compressor outlet (Ps30), align with intuitive expectations about engine performance at different flight stages. For instance, SmLPC is most vital during initial thrust buildup, while Ps30 becomes increasingly significant during descent, affecting engine stability. The Spearman correlation matrix provides further insight into these findings and highlights the limitations of relying solely on correlation matrices, which might not capture the multivariate and potentially nonlinear relationships revealed by more advanced predictive models.

In terms of future research, the focus could shift towards advanced feature engineering to capture nuanced patterns, particularly for non-nominal flights. Developing data augmentation techniques could mitigate the disparities between training and testing flights, thereby enhancing model generalizability. Temporal data trends could also be explored via specialized time-series models to track the progressive changes in engine health. Furthermore, incorporating semantic-level information could not only improve the model's predictive accuracy but also its interpretability. Given the critical safety implications of turbofan engines, the integration of uncertainty quantification methods could provide an added layer of reliability to these prognostics. Through this multifaceted analysis,

the study paves the way for more targeted, effective, and reliable turbofan engine health monitoring, offering avenues for both immediate application and future research.

REFERENCES

Sun, B., Zeng, S., Kang, R., & Pecht, M. G. (2012). Benefits and challenges of system prognostics. *IEEE Transactions on reliability*, 61(2), 323-335. doi:10.1109/TR.2012.2194173

An, D., Kim, N. H., & Choi, J. H. (2015). Practical options for selecting data-driven or physics-based prognostics algorithms with reviews. *Reliability Engineering & System Safety*, 133, 223-236. doi:10.1016/j.ress.2014.09.014

Wang, D., Miao, Q., & Kang, R. (2009). Robust health evaluation of gearbox subject to tooth failure with wavelet decomposition. *Journal of Sound and Vibration*, 324(3-5), 1141-1157. doi:10.1016/j.jsv.2009.02.013

Lee, N., Azarian, M. H. & Pecht, M. G., An explainable deep learning-based prognostic model for rotating machinery, doi:10.48550/arXiv.2004.13608

Lall, P., & Thomas, T. (2017, August). PCA and ICA based prognostic health monitoring of electronic assemblies subjected to simultaneous temperature-vibration loads. In *International Electronic Packaging Technical Conference and Exhibition* (Vol. 58097, p. V001T03A007). American Society of Mechanical Engineers. doi:10.1115/IPACK2017-74239

Dupuis, R. (2010). Application of oil debris monitoring for wind turbine gearbox prognostics and health management. In *Annual Conference of the PHM Society* (Vol. 2, No. 1). doi:10.36001/phmconf.2010.v2i1.1867

Li, X., Li J., Zuo, L., Zhu, L. & Shen, H.T. Domain adaptive remaining useful life prediction with transformer. *IEEE Transactions on Instrumentation and Measurement* 71 (2022): 1-13. doi:10.1109/TIM.2022.3200667

Biggio, L., Wieland, A., Chao, M.A., Kastani, I., Fink, O. "Uncertainty-aware prognosis via deep gaussian process." *IEEE Access 9* (2021): 123517-123527. doi:10.1109/ACCESS.2021.3110049

Sun, J., Zuo, H., Wang, W., & Pecht, M. G. (2012). Application of a state space modeling technique to system prognostics based on a health index for condition-based maintenance. *Mechanical systems and signal processing*, 28, 585-596. doi:10.1016/j.ymssp.2011.09.029

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. doi:10.1109/PHM.2008.4711414

Chao, M.A., Kulkarni, C., Goebel, K., & Fink, O. (2021). Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data,* 6(1), 5. doi:10.3390/data6010005

Frederick, D. K., DeCastro, J. A., & Litt, J. S. (2007). User's guide for the commercial modular aero-propulsion system simulation (C-MAPSS) (No. E-16205).

Srivastava, A. & Klassen, E.P., (2016) Functional Data and Elastic Registration, In *Functional and Shape Data Analysis*. Springer Series in Statistics. Springer: New York, NY, doi:10.1007/978-1-4939-4020-2.

Hong, Y. (2020). *scikit-fda: Principal Component Analysis for Functional Data.* Bachelor's thesis. Autonomous University of Madrid, Madrid, Spain.