Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira, Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26166-cd

NGAFID-Driven Whole-Airframe Health Analytics for Real-World Aircraft via Information-Propagation-Constrained Attention

Chen, Xinhang

Hangzhou International Innovation Institute, Beihang University, China. E-mail: ChenXHang@buaa.edu.cn

Wei, Zhihuan

Hangzhou International Innovation Institute, Beihang University, China. E-mail: weizhihuan@buaa.edu.cn

Baraldi, Piero

Politecnico di Milano, Italy. E-mail: piero.baraldi@polimi.it

Hu, Yang

Hangzhou International Innovation Institute, Beihang University, China. E-mail: yang_hu@buaa.edu.cn

Real-world flight profiles are high-noise, low-signal and strongly-coupled multi-task time series where weak fault signatures are easily drowned by long-range operational variations. We introduce an information-propagation-constrained attention framework that limits self-attention receptive fields to suppress cross-stage noise while retaining local fault cues. The architecture first tokenises each profile via a lightweight convolutional tokenizer (ConvTok) embedding local shape, statistics and positional codes; tokens are then processed by an AnyAttn backbone whose multi-head self-attention switches between global, sliding-window (SWLA), multi-window (MWLA) or log-parse local (LPLA) modes via learnable band masks. Experiments on the National General Aviation Flight Information Database (NGAFID)—28,000 Cessna-172 flights covering 36 maintenance classes—show that global attention achieves 79.29% anomaly-detection accuracy, while MWLA with 3-token windows raises fault classification F1 from 34.46% to 51.76%, verifying that restricted receptive fields enlarge effective signal-to-noise ratio under label noise and few-shot constraints.

Keywords: Aviation health monitoring, real-world aviation dataset, time-series classification, anomaly detection, fault classification, information-propagation constraint, local self-attention.

1. Introduction

Aviation Prognostics and Health Management (PHM) has focused on idealised component-level data (C-MAPSS (Chao et al. 2021); engine (Muneer et al. 2021; Huang et al. 2022; Feng et al. 2024) (Huang et al. 2022; Feng et al. 2024); compressors/turbine (Huang et al. 2022; Cohen, Huan, and Ni 2023)) while neglecting real-world noise, sparse sensors and few fault samples that degrade SNR and impede fault signature extraction (Saxena et al. 2008; Kumar et al. 2020; Berri, Dalla Vedova, and Mainini 2021; Huang et al. 2022). Standard baselines — Multi-Layer Perceptron (MLP) (Rumelhart and Hintont 1986), Convolutional Neural Network (CNN) (Lecun et al. 1998), Long Short-Term Memory (LSTM) (Hochreiter and Jürgen Schmidhuber 1997) — show degraded performance under high-noise, real-world conditions. Whole-airframe health monitoring remains scarce (Fu and Avdelidis 2023): heterogeneous, cross-coupled subsystems complicate fault isolation (Berri, Dalla Vedova, and Mainini 2021; Dingeldein 2023), while high-cost data collection yields persistent small-data constraints (Scott et al. 2022; Cho et al. 2023; Li et al. 2024) and sparse sensing nodes limit observability (Verhagen et al. 2023), contrasting with subsystem-focused studies (Wang et al. 2021; 2025).

Recent whole-airframe datasets exemplified by the National General Aviation Flight Information Database (NGAFID), simultaneously enable system-level research and introduce new analytical challenges (Yang and

Desell 2022). The NGAFID dataset, sourced from real flight environments of the Cessna-172 fleet, exhibits characteristics of authenticity, sparse target components, multi-task modes, and intercoupling. These characteristics increase the complexity of analysis and highlight the necessity of in-depth research on real industrial data. For example, sparse target signals and imbalanced sample sizes, low signal-to-noise ratios, and complex intercoupling between different task stages make the extraction of local fault features extremely difficult. To address these issues, several methods have been proposed. Multi-Head Self-Attention (MHSA), which computes parallel attention across multiple representation subspaces to capture diverse dependencies, has been applied in aviation PHM. Conv-MHSA (Yang, Labella, and Desell 2019) and self-supervised transfer schemes (Wang et al. 2025) mitigate data scarcity, yet they leave the core problem—global noise drowning locally weak fault signatures—largely untouched. Global-attention classifiers aggregate cross-stage variations, attenuating discriminability in real-world fault classification tasks.

Inspired by local interactions mechanism of InceptionTime (Ismail Fawaz et al. 2020), we embed information-propagation constraints into self-attention (shown in Fig. 1). Convolutional tokenizer (ConvTok) segments the profile into short tokens that encode local shape and statistics, after which AnyAttn — an MHSA backbone with learnable band masks — limits receptive fields to suppress long-range noise. Global average pooling and a linear classifier complete the pipeline; the resultant architecture significantly boosts discriminability on high-noise, real-world data.

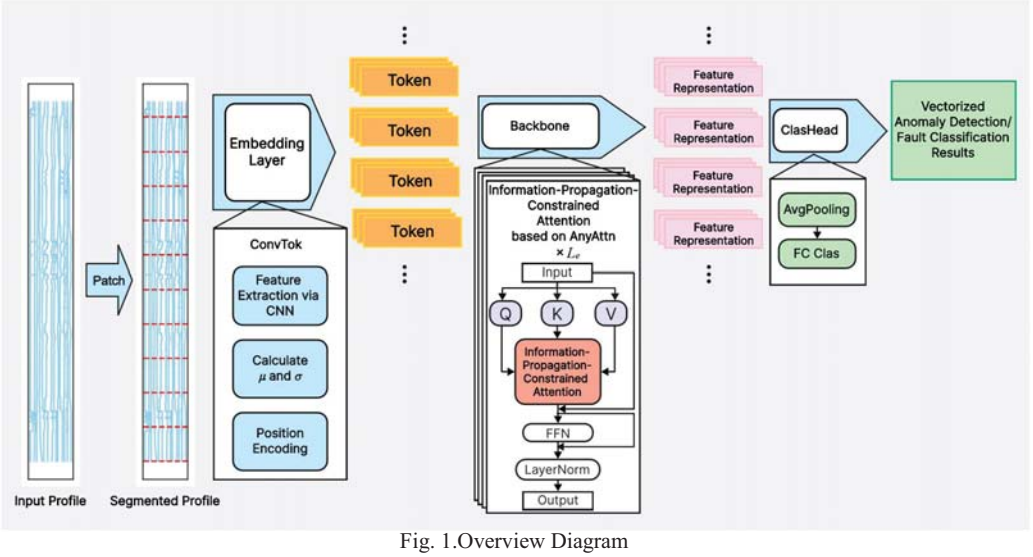

Fig. 1.Overview Diagram

2. Data Description of NGAFID Dataset

To delve into the health management of aviation equipment in real industrial scenarios, this study centers on the NGAFID aviation maintenance dataset. Derived from a fleet of Cessna-172 aircraft operated by a real flight school, the dataset encompasses over 28,000 flights and 31,000 flight hours of multivariate time-series data, correlated with more than 2,000 unscheduled maintenance events and covering 36 distinct types of maintenance issues. Unlike benchmark datasets generated in ideal simulation or laboratory settings, NGAFID is collected from a real, diverse operational environment characterized by varying seasons, weather conditions, pilot behaviors, and flight profiles, rendering it a valuable resource for studying practical PHM problems.

Each flight mission samples multivariate time-series data from 23 sensor channels at 1 Hz, yielding over 100 million records. These sensors

cover key subsystems including electrical systems, fuel systems, engine parameters, cylinder conditions, flight states, and environmental parameters (Table 1), forming a rich observational space for high-dimensional anomaly detection.

Table 1. Description of Dataset Variables

Sensor Category	Example Variable Names
Electrical System	volt1, volt2, amp1, amp2
Fuel System	FQtyL, FQtyR, E1 FFlow
Engine Parameters	E1 OilT, E1 OilP, E1 RPM
Cylinder Status	E1 CHT1-4, E1 EGT1-4
Flight Status	IAS, VSpd, AltMSL, NormAc
Environmental Parameters	OAT

The NGAFID dataset exhibits the following characteristics:

Authenticity. Real Cessna-172 operations introduce genuine noise and maintenance-record label noise.

Sparse Target Components. PCA on 20 classes (20 most representative flights per class, 20 dimensions) shows that healthy vs. abnormal separation resides in minor components 12 to 18 whose variance ratio is only around 10^{-6}. This quantitative mismatch (signal energy <0.1% of total variance) confirms that discriminative features are buried in high-dimensional, low-SNR subspaces.

Coupled multi-task modes. multi-task modes of NGAFID encapsulate the coupled nature of real Cessna-172 operations: take-off engine state propagates into cruise fuel economy and stability, while landing decisions are predicated on cruise metrics, prohibiting phase-wise isolation and mandating whole-profile modeling.

Consequently, global noise in NGAFID overwhelms local fault cues, and full-range attention amplifies this interference. Shrinking the receptive field is imperative to boost local SNR and recover weak signatures; Section 3 introduces information-propagation-constrained attention to achieve this.

3. Methodology

Motivated by the finding that the convolutional front-end of Conv-MHSA (Yang, Labella, and Desell 2019) is parameter-heavy yet SNR-sensitive, we propose the GPT-style ConvTok-AnyAttn framework (Fig. 1). ConvTok tokenises the profile to cut attention cost and enrich the input space, while the AnyAttn backbone embeds switchable, propagation-constrained self-attention that suppresses global noise and boosts both efficiency and discriminability.

3.1.*ConvTok*

Tokenization splits a length-L sequence into non-overlapping patches of length p, shrinking self-attention from $O(L^2)$ to $O(L^2/p^2)$ and furnishing a structured latent space that standard linear embeddings lack. ConvTok generates each token via CNN-extracted segment features, plus embedded statistics and positional code.

Segment Feature Encoding: A 1-D CNN encodes each length-p segment into local shape descriptors.

Token Statistics: Each token is augmented with its own mean μ_i and standard deviation σ_i, furnishing low-cost, task-stage-aware context.

Positional Encoding: Sinusoidal position code $\sin(0.1 \cdot pos)$, pos calculated as the ratio of the token index to the total number of tokens. Positional code preserves within-sequence order.

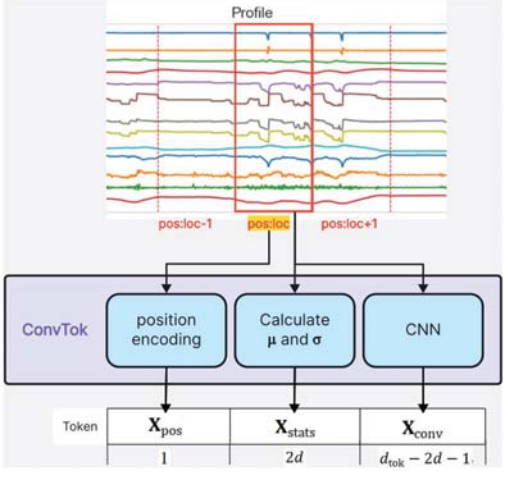

Fig. 2.Operating Mode of ConvTok

The above process can be represented as follows: Assuming the input $\mathbf{X} \in \mathbb{R}^{B\times L\times d}$, with token length p, the number of the whole tokens

$N_{\text{tok}} = \lceil L/p \rceil$, and the token dimension d_{tok} being the final feature dimension of each token, the dimension of the convolutional features $\mathbf{X}_{\text{conv}}$ is $d_{\text{conv}} = d_{\text{tok}} - 2d - 1$, the dimension of the statistical features $\mathbf{X}_{\text{stats}}$ is $2d$, and the dimension of the positional encoding $\mathbf{X}_{\text{pos}}$ is 1. See Eq. (1).

$$\begin{aligned} X_{tok} &= [X_{conv}, X_{stats}, \mathbf{X}_{\text{pos}}] \\ &= [\text{Conv1D}(\mathbf{X}; \mathbf{W}_{\text{conv}}), \mathbf{X}_{\text{stats}}, \mathbf{X}_{\text{pos}}] \end{aligned} \quad (1)$$

ConvTok fuses CNN features, token-wise statistics, and sine position codes into compact representations, cutting self-attention cost and model size while retaining rich, low-complexity features ideal for large-scale or resource-constrained deployment.

3.2.AnyAttn Framework

AnyAttn is a CNN-augmented MHSA backbone (Fig. 3), which first tokenises the profile with ConvTok for low-cost, semantics-rich segments, then replaces fully connected projections and FFN with 1-D convolutions that reinforce local pattern extraction and suppress long-range noise, yielding superior fault-classification accuracy in low-SNR data.

Each AnyAttn block keeps dimension d_{tok} constant; $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are produced by lightweight 1-D convolutions rather than fully connected projections:

$$\begin{aligned} \mathbf{Q} &= \text{Conv1D}(\mathbf{X}_{\text{front}}; \mathbf{W}_Q) \\ \mathbf{K} &= \text{Conv1D}(\mathbf{X}_{\text{front}}; \mathbf{W}_K) \\ \mathbf{V} &= \text{Conv1D}(\mathbf{X}_{\text{front}}; \mathbf{W}_V) \end{aligned} \quad (2)$$

where $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ are learnable kernel parameters and $\mathbf{X}_{\text{front}} \in \mathbb{R}^{N_{\text{tok}} \times d_{\text{tok}}}$ is the output of the preceding layer (either the embedding layer or a previous AnyAttn block). This design captures local temporal dependencies while further reducing parameter count.

The MHSA attention is computed as:

$$\mathbf{A} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \quad (3)$$

where $\mathbf{A} \in \mathbb{R}^{N_{\text{tok}} \times d_{\text{tok}}}$ is the refined feature representation and d_k is the head dimension. Section 3.3 details the available attention variants (global, SWLA, MWLA, LPLA).

The FFN also replaces fully connected layers with two successive 1-D convolutions. Gaussian Error Linear Unit (GELU) is employed for superior nonlinear expressiveness and training stability:

$$\begin{aligned} &\text{FFN}(\mathbf{A}) = \\ &\text{Conv1D}(\text{GELU}(\text{Conv1D}(\mathbf{A}; \mathbf{W}_1)); \mathbf{W}_2) \end{aligned} \quad (4)$$

Residual connections followed by layer normalization stabilize training:

$$\mathbf{X}_{\text{out}} = \text{LN}(\mathbf{X}_{\text{front}} + \text{SubLayer}(\mathbf{X}_{\text{front}})) \quad (5)$$

where SubLayer denotes either the self-attention or the FFN module. Thus, every AnyAttn block contains two residual shortcuts.

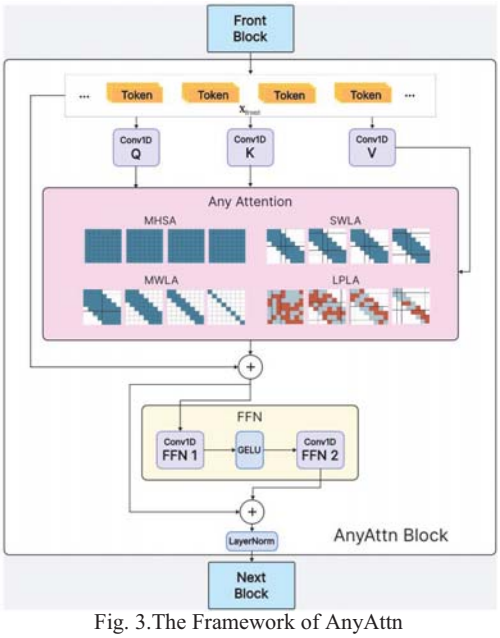

Fig. 3.The Framework of AnyAttn

Let L_e denote the number of stacked AnyAttn blocks. After the embedding layer and L_e AnyAttn encodings, the sequence is globally average-pooled along the temporal dimension and fed to a single unbiased linear classifier:

$$\hat{\mathbf{y}} = \text{Linear}\left(\frac{1}{N_{\text{tok}}}\sum_{i=1}^{N_{\text{tok}}} \mathbf{X}_{\text{aao}_i}\right) \quad (6)$$

where $\mathbf{X}_{\text{aao}_i}$ is the final latent vector of the i-th token, and the output dimension equals the number of classes for the target task.

This enhanced framework effectively handles temporal characteristics and provides a solid foundation for subsequent attention-mechanism refinements.

3.3.*Information-Propagation-Constrained Attention*

Global self-attention aggregates coupled-stage dependencies but simultaneously injects task-irrelevant noise, degrading the low SNR of industrial sequences. We therefore constrain propagation via learnable binary masks **M** that zero attention outside a prescribed region, yielding three variants — SWLA, MWLA and LPLA (Fig. 3). To restrict information flow we insert **M** in attention process:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T \odot \mathbf{M}}{\sqrt{d_k}}\right)\mathbf{V} \quad (7)$$

with $\odot$ denoting element-wise multiplication. In multi-head form the input is split into H heads, each endowed with an independent mask $\mathbf{M}_i$:

$$\text{MH}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_H)\mathbf{W}^O \quad (8)$$

$$\text{head}_i = \text{softmax}\left(\frac{(\mathbf{Q}\mathbf{W}_i^Q)(\mathbf{K}\mathbf{W}_i^K)^T \odot \mathbf{M}_i}{\sqrt{d_k}}\right)(\mathbf{V}\mathbf{W}_i^V) \quad (9)$$

where $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, $\mathbf{W}_i^V$ are head-specific projection kernels. Below we detail how each variant defines $\mathbf{M}_i$.

SWLA enforces a single fixed-width sliding window of half-size w:

$$\mathbf{M}_{\text{SWLA}}(i,j) = \begin{cases} 1, & if\ |i-j| \le w \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

Only tokens within w positions can interact, forcing the model to focus on local neighborhoods and ignore distant disturbances.

MWLA extends SWLA by equipping each head with its own half-width w_i, so one layer concurrently extracts fine, medium and coarse patterns without extra depth or width:

$$\mathbf{M}_{\text{MWLA},i}(j,k) = \begin{cases} 1, & if\ |j-k| \le w_i \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

LPLA hybridizes locality and globality: most heads use narrow windows while one or more set $w_i = 0$ for long-range context; LogParse sparsification retains only top attention values, blocking noise re-introduction.

$$\mathbf{M}_{\text{LPLA},i}(j,k) = \begin{cases} 1, & if\ |j-k| \le w_i \text{ or } w_i = 0 \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

SWLA, MWLA and LPLA systematically compress the receptive field, boosting local SNR to expose weak fault signatures within complex operational profiles; empirical validation is given in Section 4.

4. Experimental Results and Discussion

4.1.*Task Definition*

Anomaly Detection: A binary classifier outputs $\mathbf{y}_{\text{detec}} = [y_{\text{healthy}}, y_{\text{abnormal}}]$, where flights are labeled abnormal if they occur within the five sorties preceding an unscheduled maintenance event, and healthy otherwise.

Fault Classification: An N_{class}-class classifier outputs $\mathbf{y}_{\text{fault}} = [y_1, ..., y_{N_{\text{class}}}]$, with each abnormal flight assigned to a specific maintenance logbook category and trained using one-hot encoding.

4.2.*Data Processing*

We use the 11,446-flight "2-day" NGAFID subset (Yang and Desell 2022). Missing data are forward-filled; sequences shorter than 900 pts or channels with >10% missingness are dropped. Profiles are cubic-spline resampled to 2,048 pts. Stratified five-fold cross-validation is enforced: records of different folds reside in separate files to avert leakage, and channel-wise z-score statistics are computed only on the training subset and then frozen. Anomaly labels are balanced; for fault classification the 5,298 abnormal flights follow a long-tail distribution and are rebalanced via controlled replication $N_{c_{\text{da}}} = \min(k_{\text{da}} \times N_c, N_{c_{\max}})$ without generating synthetic samples for data augmentation (DA).

4.3.*Experimental Setup*

Each of the five stratified folds is trained de novo; no cross-fold weight or hyper-parameter reuse occurs to preclude information bleed. Early stopping (patience=3) and an adaptive learning rate (halved on plateau, minimum 5×10^{-12}) counteract over-fit. Adam (β_1=0.9, β_2=0.999) minimizes categorical cross-entropy; fold-wise test metrics — Accuracy, Recall and F1 — are computed from the concatenated predictions of the held-out set. Code: Python 3.12+PyTorch 2.7.1; hardware: i7-13620H, 32 GB RAM, RTX-4070 28 GB VRAM.

4.4.*Comparative Results*

Comprehensive comparisons were performed for both tasks, with methods optimized individually using distinct hyperparameters.

For anomaly detection, The baseline models MLP, LSTM, CNN and InceptionTime set d_{hidden}=128; InceptionTime sets 4 layers, d_{hidden}=2048, filters=128 and initial learning rate=3×10^{-5}. ConvTok-based models share p=4, CNN kernel size = 3, 2 layers, 2 heads, d_{FFN}=512, d_{tok}=128 and initial learning rate=1×10^{-4}; SWLA uses w=5, MWLA uses w_i in [3, 5], and LPLA uses w_i in [0, 5] (heads with 0 revert to global). Batch sizes are 256 for MLP/CNN, 64 for LSTM, 32 for InceptionTime and all ConvTok variants.

For fault classification, depths and widths are increased where needed: InceptionTime employs 6 layers, d_{hidden}=2048, filters=256; ConvTok models use 4 layers, 4 heads, d_{FFN}=1024, d_{tok}=512. SWLA sets w=3, MWLA w_i in [0, 1, 3, 5], and LPLA w_i in [0, 1, 3, 5] (same with anomaly detection, heads with 0 revert to global). Initial learning rates and batch sizes remain 1×10^{-4} and 32, respectively, for all ConvTok variants; other baselines keep the same settings as their anomaly-detection counterparts. Hyperparameters were determined via manual grid search on the validation set, prioritizing convergence stability over marginal gains.

Table 2. Anomaly Detection Experimental Results

Model	Accuracy	Recall	F1
MLP	0.6323	0.6323	0.6323
LSTM	0.5664	0.5676	0.5647
CNN	0.6495	0.6508	0.6478
InceptionTime	0.7669	0.7667	0.7668
ConvTokMHSA	**0.7929**	**0.7931**	**0.7929**
ConvTokSWLA	0.7660	0.7655	0.7655
ConvTokMWLA	0.7591	0.7506	0.7589
ConvTokLPLA	0.7841	0.7838	0.7839

Anomaly Detection. Table 2 shows global-receptive ConvTok-MHSA and ConvTok-LPLA achieve highest accuracy (79.29%/78.41%), confirming whole-profile context is indispensable.

Table 3. Fault Classification Experimental Results

Model	Metrics	Without DA	With DA (k_{da}=2)
MLP	Accuracy	0.3660	0.3094
	Recall	0.0558	0.0782
	F1 Score	0.0364	0.0789
CNN	Accuracy	0.3566	0.3057
	Recall	0.0546	0.0601
	F1 Score	0.0357	0.0472
LSTM	Accuracy	0.3623	0.2745
	Recall	0.0526	0.0572
	F1 Score	0.0280	0.0373
Inception Time	Accuracy	0.5264	0.5726
	Recall	0.3487	**0.4979**
	F1 Score	0.3950	0.5067
ConvTok MHSA	Accuracy	0.4094	0.4255
	Recall	0.1963	0.3343
	F1 Score	0.2183	0.3446
ConvTok SWLA	Accuracy	0.5208	0.5660
	Recall	0.3790	0.4744
	F1 Score	0.4126	0.5030
ConvTok MWLA	Accuracy	**0.5462**	**0.5840**
	Recall	**0.4020**	0.4694
	F1 Score	**0.4310**	**0.5176**
ConvTok LPLA	Accuracy	0.3708	0.4349
	Recall	0.1896	0.2563
	F1 Score	0.2013	0.2919

Fault Classification. Table 3 shows strictly local ConvTok-SWLA/MWLA maximize F1; MWLA attains 51.76% (+17.3 pp over ConvTok-MHSA), corroborating short-duration cues are corrupted by long-range context. Hybrid LPLA underperforms, confirming limited global reach degrades temporal precision.

ConvTok-based models achieve lightweight deployment: ConvTokMHSA processes 512 tokens at ~7.8s/epoch with 1.54MB size and 0.01s inference (32 samples); restricted variants (SWLA, MWLA, LPLA) maintain 1.54MB with comparable inference (0.01–0.04s), trading modest per-epoch overhead (10–34s) for improved FC accuracy.

4.5.*Receptive-Field Analysis*

The receptive field (RF) quantifies the length of the raw profile that influences a single temporal prediction. Under ConvTokSWLA, RF depends only on the token size p and the window half-width w:

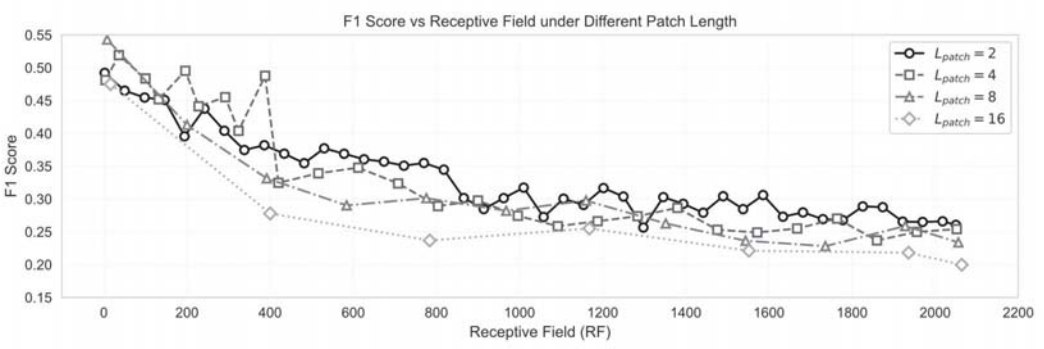

Fig. 4.Impact of Different Segment Lengths on F1-Score under Receptive Field

$$\mathrm{RF} = (2 \times w \times L_e + 1) \times p \qquad (13)$$

Because L_e is the number of non-overlapping AnyAttn layers, RF grows linearly with w and L_e.

Fig. 4 shows that Fault Classification F1 decreases monotonically as RF increases. A small RF forces the model to focus on neighborhood patterns and suppresses cross-stage noise, whereas an RF that is too large dilutes fault signatures with irrelevant long-range variations. At equal RF, models with finer segmentation (p=2 to 8) consistently outperform those with p=16, confirming that fine-grained local modeling and information-propagation constraints act synergistically.

5. Conclusion and Future Work

We deliver ConvTok-AnyAttn, an information-propagation-constrained pipeline; global context yields 79.29% anomaly-detection accuracy, while localized attention achieves 51.76% fault-classification F1 on the noisy NGAFID set, corroborating that receptive-field restriction recovers short-duration fault signatures masked by operational drift.

Current limitations are three-fold.

(i) The present token grid omits subsystem-interaction topology; Graph Neural Network is planned.

(ii) Sliding-window attention incurs quadratic memory, constraining ultra-long sorties; Flash Attention or equivalent linear-complexity variants will be adopted.

(iii) Uniform channel-noise modeling is sub-optimal; learnable channel-gating will next suppress low-informative sensors to raise effective SNR.

Overall, the study offers a deployable framework for whole-airframe PHM under scarce, noisy labels; ongoing work targets architectural refinement, memory-efficient attention, and adaptive denoising to accelerate fleet-wide industrial adoption.

References

Berri, Pier Carlo, Matteo D.L. Dalla Vedova, and Laura Mainini. 2021. "Computational Framework for Real-Time Diagnostics and Prognostics of Aircraft Actuation Systems." *Computers in Industry* 132:103523. https://doi.org/10.1016/j.compind.2021.103523.

Chao, M., C. Kulkarni, K. Goebel, and O. Fink. 2021. "Aircraft Engine Run-To-Failure Dataset Under Real Flight Conditions." *NASA Ames Prognostics Data Repository*. https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/#turbofan-2.

Cho, Seong Hee, Seokgoo Kim, Jung Heon Lee, Jinwoo Song, and Joo Ho Choi. 2023. "A Comparative Study of Data-Driven Prognostic Approaches under Training Data Deficiency." *IFAC-PapersOnLine* 56 (2): 3744–50. https://doi.org/10.1016/j.ifacol.2023.10.1543.

Cohen, Joseph, Xun Huan, and Jun Ni. 2023. "Fault Prognosis of Turbofan Engines: Eventual Failure Prediction and Remaining Useful Life Estimation." *International Journal of Prognostics and Health Management* 14 (2): 1–10. https://doi.org/10.36001/IJPHM.2023.v14i2.3486.

Dingeldein, Lorenz. 2023. "Simulation Framework for Real-Time PHM Applications in a System-of-Systems Environment." *Aerospace* 10 (1). https://doi.org/10.3390/aerospace10010058.

Feng, Kun, Yuan Xiao, Zhouzheng Li, and Dongyan Miao. 2024. "A Fusion Autoencoder Model and Piecewise Anomaly Index for Aero-Engine Fault Diagnosis." *Applied Intelligence* 54 (20): 10148–60. https://doi.org/10.1007/s10489-024-05712-7.

Fu, Shuai, and Nicolas P. Avdelidis. 2023. "Prognostic and Health Management of Critical Aircraft Systems and Components: An Overview." *Sensors* 23 (19). https://doi.org/10.3390/s23198124.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9 (8): 1735–80.

Huang, Yufeng, Jun Tao, Gang Sun, Hao Zhang, and Yan Hu. 2022. "A Prognostic and Health Management Framework for Aero-Engines Based on a Dynamic Probability Model and LSTM Network." *Aerospace* 9 (6). https://doi.org/10.3390/aerospace9060316.

Ismail Fawaz, Hassan, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre Alain Muller, and François Petitjean. 2020. "InceptionTime: Finding AlexNet for Time Series Classification." *Data Mining and Knowledge Discovery* 34 (6): 1936–62. https://doi.org/10.1007/s10618-020-00710-y.

Kumar, Anil, C. P. Gandhi, Yuqing Zhou, Rajesh Kumar, and Jiawei Xiang. 2020. "Latest Developments in Gear Defect Diagnosis and Prognosis: A Review." *Measurement: Journal of the International Measurement Confederation* 158:107735. https://doi.org/10.1016/j.measurement.2020.107735.

Lecun, Yann, Leon Bottou, Yoshua Bengio, and Patrick Ha. 1998. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86 (11): 2278–2324.

Li, Chuanjiang, Shaobo Li, Yixiong Feng, Konstantinos Gryllias, Fengshou Gu, and Michael Pecht. 2024. *Small Data Challenges for Intelligent Prognostics and Health Management: A Review*. *Artificial Intelligence Review*. Vol. 57. Springer Netherlands. https://doi.org/10.1007/s10462-024-10820-4.

Muneer, Amgad, Shakirah Mohd Taib, Sheraz Naseer, Rao Faizan Ali, and Izzatdin Abdul Aziz. 2021. "Data-Driven Deep Learning-Based Attention Mechanism for Remaining Useful Life Prediction: Case Study Application to Turbofan Engine Analysis." *Electronics (Switzerland)* 10 (20). https://doi.org/10.3390/electronics10202453.

Rumelhart, David E, and Geoffrey E Hintont. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323 (6088): 533–36.

Saxena, Abhinav, Kai Goebel, Don Simon, and Neil Eklund. 2008. "Damage Propagation Modeling for Aircraft Engine Run-to-Failure Simulation." *2008 International Conference on Prognostics and Health Management, PHM 2008*. https://doi.org/10.1109/PHM.2008.4711414.

Scott, Michael J, Wim J C Verhagen, Marie T Bieber, and Pier Marzocca. 2022. "A Systematic Literature Review of Predictive Maintenance for Defence Fixed-Wing Aircraft Sustainment and Operations." *Sensors* 22 (18): 7070.

Verhagen, Wim J.C., Bruno F. Santos, Floris Freeman, Paul van Kessel, Dimitrios Zarouchas, Theodoros Loutas, Richard C.K. Yeun, and Iryna Heiets. 2023. "Condition-Based Maintenance in Aviation: Challenges and Opportunities." *Aerospace* 10 (9): 1–23. https://doi.org/10.3390/aerospace10090762.

Wang, Yilin, Peixuan Lei, Xuyang Wang, Liangliang Jiang, Liming Xuan, Wei Cheng, Honghua Zhao, and Yuanxiang Li. 2025. "Leveraging Large Self-Supervised Time-Series Models for Transferable Diagnosis in Cross-Aircraft Type Bleed Air System." *Advanced Engineering Informatics* 65 (PC): 103275. https://doi.org/10.1016/j.aei.2025.103275.

Wang, Yilin, Yifei Yu, Kong Sun, Peixuan Lei, Yuxuan Zhang, Enrico Zio, Aiguo Xia, and Yuanxiang Li. 2021. "RmGPT: A Foundation Model With Generative Pre-Trained Transformer for Fault Diagnosis and Prognosis in Rotating Machinery." *IEEE Internet of Things Journal* 14 (8): 1–12. https://doi.org/10.1109/JIOT.2025.3580823.

Yang, Hong, and Travis Desell. 2022. "A Large-Scale Annotated Multivariate Time Series Aviation Maintenance Dataset from the NGAFID," no. 2012. http://arxiv.org/abs/2210.07317.

Yang, Hong, Aidan Labella, and Travis Desell. 2019. "Predictive Maintenance for General Aviation Using Convolutional Transformers."