

Artificial Intelligence for Aircraft Fault Detection: From Accuracy to Airworthiness through Modality-Aware Analysis, Validation, and Certification Perspectives

Rexcharles Enyinna Donatus¹; Loveth O. Uhiah¹; Christian Onyewuchi Njoku²; and Ifeyinwa Happiness Donatus³

¹Aerospace Engineering Department, Faculty of Air Engineering, Air Force Institute of Technology, Kaduna, Nigeria. ²Mechanical Engineering Department, Faculty of Air Engineering, Air Force Institute of Technology, Kaduna, Nigeria. ³Department of Computer Science, Kaduna State University, Kaduna, Nigeria.

Corresponding author: charlly4eyims@yahoo.com

DOI: <https://doi.org/10.62154/ajastr.2026.022.01019>

Abstract

Artificial Intelligence (AI)-based fault detection has demonstrated high diagnostic accuracy in aircraft systems, with many studies reporting performance exceeding 90–99%. However, these results are predominantly obtained under controlled, offline, or simulation-based conditions, limiting their relevance for safety-critical aviation environments where failure consequences are severe and certification requirements are stringent. This gap highlights the need for evaluation frameworks that extend beyond accuracy to operational and airworthiness considerations. This survey presents a modality-aware analytical framework for AI-based aircraft fault detection, organizing existing methods along four key dimensions: input modality, data provenance, validation maturity, and model class. Using this taxonomy, this survey systematically reviews signal, image, log, and simulation-based approaches, demonstrating that reported performance is strongly influenced by data characteristics and evaluation conditions. The analysis shows that many high-performing models rely on low-maturity validation and lack evidence of generalization under real-world operating conditions. To address these limitations, this survey introduces an engineering-oriented deployment framework integrating modality-driven model selection, structured data preprocessing, multi-level validation, computational feasibility, and explainability. A deployment readiness model is further proposed to assess systems based on validation maturity, data coverage, interpretability, and operational constraints. This survey identifies key challenges, including the simulation-to-reality performance gap, lack of standardized benchmarks, limited adoption of explainable AI, and insufficient alignment with certification requirements, and outlines future research directions toward certification-aware AI design, lightweight deployable models, and multimodal digital twin integration.

Keywords: Artificial Intelligence, Aircraft Fault Detection, Predictive Maintenance, Validation Maturity, Explainable AI, Aerospace Certification.

Introduction

Ensuring timely and reliable fault detection in aircraft systems is fundamental to flight safety, dispatch reliability, and maintenance efficiency. Modern aircraft operate as highly

instrumented systems, generating high-frequency data from engines, fuel systems, actuators, sensors, and structural components. At the same time, operators face increasing pressure to reduce unplanned maintenance events and improve operational availability (Kabashkin et al., 2025; Kwakye et al., 2024). Traditional maintenance strategies based on scheduled inspections and conservative replacement intervals are becoming insufficient in the context of growing system complexity and variable operational conditions (Lei et al., 2024; M. Li et al., 2025). The proliferation of onboard sensing, communication systems, and maintenance data infrastructures has therefore created significant opportunities for the adoption of AI-driven approaches to fault detection and condition-based maintenance (Kabashkin, 2025; Kwakye et al., 2024; Lei et al., 2024; J. Li et al., 2023).

Historically, aircraft fault detection has relied on rule-based diagnostics, threshold-based monitoring, and physics-based models tailored to specific subsystems. These approaches provide interpretability and alignment with certification requirements but are limited in handling nonlinear system behavior, complex interactions, and high-dimensional data (Ezhilarasu et al., 2021; Lei et al., 2024; Zolghadri, 2024). Early machine learning methods introduced data-driven capabilities through supervised classifiers and statistical models, but their dependence on handcrafted features and limited scalability restricted their effectiveness in complex operational environments (Ezhilarasu et al., 2021).

Recent advances in artificial intelligence, particularly deep learning, have significantly enhanced fault detection capabilities. Architectures such as convolutional neural networks, recurrent neural networks including long short-term memory models, and hybrid CNN–LSTM frameworks enable automated feature extraction from large-scale sensor data and support improved anomaly detection across diverse aircraft subsystems. These methods have demonstrated high diagnostic accuracy in applications such as vibration analysis, engine health monitoring, and fleet-level data analytics (Du et al., 2024; Kabashkin, 2025; Khanzada et al., 2025). In parallel, AI-enabled maintenance platforms are increasingly integrating such models into predictive maintenance workflows and decision-support systems (Kabashkin et al., 2025; Kwakye et al., 2024).

Despite these advances, several challenges continue to limit the practical deployment of AI-based fault detection systems. Data scarcity and imbalance remain significant issues, particularly for rare but safety-critical faults in fleet-scale datasets (Dangut et al., 2023). Domain shift between simulated, rig-based, and real operational data affects model generalization and robustness (Ezhilarasu et al., 2021; Kraemer & Villani, 2024; J. Li et al., 2025).

Interpretability is another critical concern, as complex models often fail to provide transparent and traceable diagnostic outputs required for engineering decision-making and certification (J. Li et al., 2023; Zolghadri, 2024). In addition, computational constraints limit the deployment of high-capacity models in onboard systems, particularly where real-time processing is required (Kabashkin, 2025; Shafi et al., 2023). Certification challenges further complicate adoption, as regulatory frameworks for data-driven systems are still evolving

and require robust validation, traceability, and reliability evidence (Luettig et al., 2024; Sridhar et al., 2025; Zolghadri, 2024) .

Recent review literature reflects both the progress and the limitations in this field. Studies on aircraft subsystems and maintenance platforms highlight the effectiveness of AI-based diagnostics while emphasizing the need for improved interpretability, data quality, and system integration (Ezhilarasu et al., 2021; Kwakye et al., 2024; J. Li et al., 2023). Reviews focused on aero engines and structural systems similarly report strong performance of deep learning approaches but stress the importance of validation under realistic operating conditions (Lei et al., 2024). Broader surveys in aviation and infrastructure domains identify common challenges related to scalability, governance, and deployment in safety-critical systems (Ahmed, 2025; Azanaw, 2024; Plevris & Papazafeiropoulos, 2024). At the same time, emerging work on certification highlights the gap between laboratory-scale model performance and the evidence required for airworthiness approval (Luettig et al., 2024; Zolghadri, 2024).

However, existing reviews are typically organized around specific application domains rather than providing a unified perspective across the diverse data modalities used in aircraft fault detection (Azanaw, 2024; Kwakye et al., 2024; Tsallis et al., 2025). Signal-based, image-based, log-based, and simulation-driven approaches differ substantially in data characteristics, preprocessing requirements, model design, and deployment feasibility. These differences are rarely integrated into a single analytical framework. Furthermore, variations in data provenance, ranging from simulation and rig experiments to real-flight and fleet-level data, and differences in validation maturity are often not explicitly distinguished (Dangut et al., 2023; Kabashkin, 2024; Kraemer & Villani, 2024; Wu et al., 2024). As a result, assessing operational readiness and certification potential remains difficult.

In response to the identified gaps, this survey makes the following contributions:

(a) Modality-aware taxonomy

A unified taxonomy for AI-based aircraft fault detection is introduced, organizing existing methods according to input modality, data provenance, validation maturity, and model class. This structure provides a consistent basis for comparing heterogeneous approaches and clarifies how modality-specific characteristics influence model design, performance, and certification relevance.

(b) Cross-modal analytical synthesis

A systematic cross-modal analysis is conducted across signal, image, log, and simulation-based approaches. The survey identifies performance trends, robustness limitations, and key trade-offs between accuracy, interpretability, and deployability, demonstrating how reported results depend on modality and data conditions (Ezhilarasu et al., 2021; Kraemer & Villani, 2024; Y. Zhao et al., 2022) .

(c) Validation and certification perspective

Validation maturity levels are explicitly characterized and linked to certification requirements in safety-critical aviation systems. The analysis highlights the gap

between high offline performance and the evidence required for airworthiness, emphasizing the need for cross-condition, transfer, and in-service validation to support trustworthy deployment (Kabashkin et al., 2025; Luettig et al., 2024; Sridhar et al., 2025; Zolghadri, 2024).

(d) Deployment readiness framework

An engineering-oriented framework is proposed to assess AI-based fault detection systems based on modality–model alignment, data provenance, validation maturity, computational feasibility, and interpretability. This framework provides practical guidance for transitioning from experimental models to deployable, certification-aware solutions in aircraft maintenance and health management.

The remainder of this survey is organized as follows. Section 2 introduces the modality-aware taxonomy and classification framework. Section 3 reviews AI-driven fault detection methods across modalities and model classes. Section 4 presents cross-modal analysis and discusses key trade-offs and limitations. Section 5 translates these findings into engineering implications and proposes a deployment readiness framework. Section 6 outlines future research directions, and Section 7 concludes the survey.

Survey Framework and Analytical Taxonomy

Purpose of the Framework

A structured analytical framework is necessary because AI-driven fault detection in aircraft systems operates across heterogeneous data modalities, distinct preprocessing pipelines, and varying levels of operational validation. Treating these diverse approaches as a single workflow obscures key engineering trade-offs, particularly those related to robustness, interpretability, and deployment feasibility. For example, time-series sensor models depend heavily on temporal alignment and sequence modeling, whereas image-based approaches are sensitive to transformation and feature encoding choices (Z. Jia et al., 2023; Wu et al., 2024).

Equally critical is the distinction between simulation-based evaluations and real-flight or fleet-level validation. High performance on benchmark datasets does not necessarily translate to operational reliability due to domain shift, noise, and incomplete labeling in real environments (Huang et al., 2023; Zolghadri, 2024). For safety-critical aerospace applications, these differences directly affect certification and trust.

Accordingly, this survey introduces a modality-aware analytical taxonomy that organizes AI-based aircraft fault detection methods along four key axes: input modality, data provenance, validation maturity, and model class. These dimensions are selected because they directly govern diagnostic performance, robustness, and certification feasibility. By explicitly structuring the literature along these axes, the proposed framework enables consistent cross-study comparison and provides a foundation for assessing deployment readiness and airworthiness relevance.

To structure the diverse body of literature on AI-based aircraft fault detection, figure 1 visualizes this framework and positions existing studies within this multi-dimensional space.

This taxonomy provides the foundation for the subsequent analysis, enabling a systematic comparison of methods beyond isolated performance metrics.

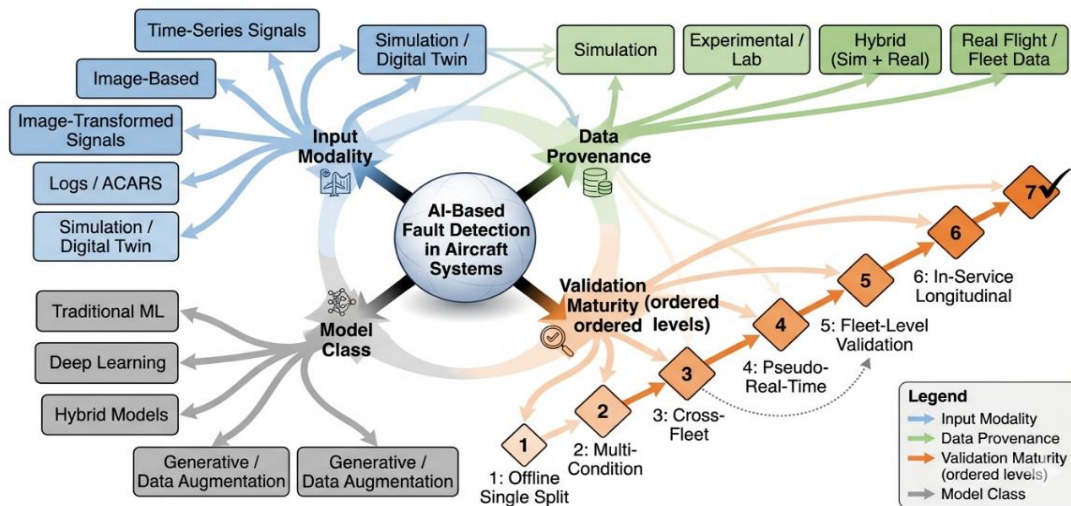


Figure 1: Taxonomy of AI- Based Aircraft Fault Detection

Literature Identification Strategy

This survey adopts a structured but non-systematic literature review approach intended to support transparent and reproducible synthesis without aiming for PRISMA-level exhaustiveness. Elements of systematic literature review practice were incorporated to improve consistency in study identification, screening, and comparative analysis.

Literature was collected from major scientific databases widely used in aerospace engineering, artificial intelligence, and predictive maintenance research, including IEEE Xplore, Scopus, Semantic Scholar, and Google Scholar. The search covered publications from 2020 to early 2026 and used combinations of keywords such as "aircraft fault detection," "predictive maintenance," "aircraft diagnostics," "deep learning," "machine learning," "structural health monitoring," "digital twin," and "aviation AI."

Inclusion was guided by three criteria: (i) clear aerospace application, (ii) explicit use of AI or hybrid diagnostic methods, and (iii) documented data source and validation procedure. Studies focused solely on non-aerospace systems, lacking empirical validation, or without clear data provenance were excluded. Screening was conducted through title, abstract, and full-text assessment to ensure methodological relevance and alignment with the survey objectives.

The search process initially identified a broad collection of studies across multiple modalities and application domains. After duplicate removal and relevance screening, more than 110 representative studies were retained for qualitative synthesis and cross-modal analysis. The objective of the search strategy was to capture methodological diversity and

deployment-relevant evidence across modalities rather than achieve exhaustive statistical coverage. The resulting corpus spans signal-based, image-based, log-based, and simulation-driven fault detection approaches with varying levels of operational validation and deployment maturity.

Conceptual papers and review studies were included selectively to support discussions on certification, deployment readiness, and system-level health management. This approach provides sufficient methodological transparency while preserving the flexibility required for a modality-aware engineering survey rather than a strictly systematic review. The final corpus included 49 representative studies spanning signal-based, image-based, log-based, and simulation-driven fault detection approaches.

Core Taxonomy

The proposed taxonomy is defined along four orthogonal axes: input modality, data provenance, validation maturity, and model class. These axes collectively capture the key technical and operational factors that determine whether an AI-based fault detection system can transition from experimental validation to certification-ready deployment.

Input Modality

Four primary modalities are identified:

- Signal (time-series data): vibration, pressure, flow, and engine parameters
- Image: raw images and transformed representations such as spectrograms
- Log and maintenance data: ACARS messages and fleet-level records
- Simulation and digital twin data: physics-based and synthetic datasets

Modality determines preprocessing requirements, feature representation, and model suitability. Signal-based methods must address temporal dependencies and noise, while image-based approaches depend on transformation fidelity. Log data introduces sparsity and class imbalance, and simulation-based data raises concerns regarding representativeness and domain shift (Huang et al., 2023; Kabashkin et al., 2025).

From an engineering perspective, modality also affects interpretability and traceability, which are critical for certification.

Data Provenance

Data provenance is categorized into four types: simulation, experimental rigs, hybrid datasets, and real-flight or in-service data. This distinction is essential because it defines the credibility of reported performance.

Simulation-based studies support model development but lack operational variability, while real-flight data captures realistic noise, imbalance, and uncertainty. Hybrid and rig-based datasets provide intermediate fidelity. Explicit separation of these categories prevents misleading comparisons and aligns evaluation with certification expectations (Huang et al., 2023).

Validation Maturity

Validation maturity is structured into four levels:

1. Offline single-scenario evaluation
2. Cross-condition validation
3. Transfer learning and domain shift analysis
4. In-service or pseudo real-time validation

This axis reflects the progression from controlled experiments to operational deployment. Higher maturity levels provide stronger evidence of robustness and reliability under real-world conditions, which is essential for safety-critical adoption (J. Li et al., 2023; Zolghadri, 2024).

Certification standards such as DO-178C for airborne software, DO-254 for airborne electronic hardware, and emerging guidance including the EASA AI Roadmap and ED-324 provide the regulatory context for assessing validation maturity and deployment readiness of AI-based fault detection systems. Traditional certification frameworks emphasize traceability, verification rigor, and deterministic behavior, whereas AI models introduce probabilistic and data-driven decision processes that require additional assurance measures related to robustness, explainability, and dataset representativeness (Dmitriev et al., 2021; Luettig et al., 2024). Consequently, models validated only under simulation or offline conditions provide insufficient evidence for safety-critical deployment. Recent certification-oriented initiatives increasingly advocate lifecycle-based "learning assurance" approaches that incorporate data governance, domain-shift evaluation, explainability, and operational robustness across the entire AI lifecycle (Durand et al., 2023; Kaakai et al., 2022). Within the proposed taxonomy, higher validation maturity levels therefore correspond not only to improved operational realism but also to stronger alignment with certification expectations, particularly for applications involving engine health monitoring, flight-control diagnostics, and other safety-critical aircraft systems.

Model Class

Model classes are grouped into: traditional machine learning, deep learning, hybrid approaches, physics-informed methods, and generative models.

This classification captures the trade-off between accuracy, interpretability, and computational cost. Deep learning models typically achieve high performance but may lack transparency and require significant resources, whereas traditional and physics-informed methods offer greater interpretability and alignment with engineering principles (Brunton et al., 2021; J. Li et al., 2023).

Analytical Mapping Strategy

The taxonomy supports a two-stage analysis. First, each study is coded across the four axes, enabling structured comparison within and across modalities. Second, cross-modal synthesis is performed by examining performance metrics in relation to data provenance and validation maturity.

Where quantitative comparison is not feasible, qualitative patterns are extracted, including trade-offs between accuracy and interpretability, and between model complexity and deployability. This mapping also highlights underexplored areas, such as models evaluated only under low validation maturity or modalities lacking real-world validation.

Limitations of the Framework

The framework is subject to several limitations. The literature coverage is selective and may reflect publication bias toward well-documented or successful models. Performance comparisons are constrained by inconsistent metrics and datasets across studies.

In addition, factors such as human–machine interaction, operational integration, and economic constraints are not explicitly modeled, although they influence deployment decisions. These limitations are considered in the interpretation of results and in the formulation of research directions. In addition, some analytical sections rely heavily on a limited number of recent fleet-level and system-integration studies, particularly (Kabashkin, 2025) and (Kabashkin et al., 2025), due to the relatively limited availability of operational-scale AI maintenance studies in the current literature. Although these works provide valuable insights into deployment and maintenance integration, this concentration may introduce evidence imbalance and reflects the broader scarcity of publicly documented large-scale aviation AI implementations. Building on this taxonomy, the next section 3 systematically reviews AI-based fault detection methods across modalities and model classes.

Survey of AI-Based Fault Detection Methods in Aircraft Systems

This section presents a structured survey of AI-based fault detection methods in aircraft systems, organized along input modality, data provenance, validation maturity, and model class. The focus is on reporting representative methods, datasets, and performance outcomes without cross-cutting interpretation, which is deferred to Section 4.

Input Modality: Signal, Image, Log, and Simulation-Based Approaches

Time-series sensor data constitute the most widely used modality in aircraft fault detection. Studies focusing on engines, actuators, and air data systems commonly employ convolutional and recurrent architectures to capture temporal dependencies. CNN–LSTM and related hybrid models applied to multi-channel sensor streams typically report classification accuracies in the range of 90% to 97%, with F1-scores approaching 0.97 under controlled experimental settings (Kabashkin, 2025; Shen & Khorasani, 2020; Y. Zhao et al., 2022). In rotor vibration analysis, pure CNN architectures operating directly on high-dimensional time-series inputs have achieved comparable performance, with reported accuracies of approximately 97% (Young et al., 2025).

A related approach transforms time-series signals into image-like representations, such as time–frequency maps or residual images, enabling the application of two-dimensional CNNs and transfer learning (R. E. Donatus, 2025). Methods based on multi-domain signal

transformations and fusion architectures report improved diagnostic accuracy relative to single-representation baselines (Z. Jia et al., 2023). Transfer learning with architectures such as VGG16, combined with channel pruning and fine-tuning, has demonstrated accuracies exceeding 98% across multiple aircraft and flight conditions, including validation on real flight data (L. Jia et al., 2025; Z. Li et al., 2023).

Image-based methods are also applied directly to visual inspection tasks. Using aerial or ground-based imagery, object detection models such as YOLOv8 have been employed to identify structural defects, including corrosion and missing components, achieving mean average precision values around 85% (Connolly et al., 2024; Liu et al., 2022).

Log and maintenance data represent another important modality, particularly for fleet-level fault detection. These datasets are characterized by sparsity and class imbalance. Deep architectures combining autoencoders and recurrent networks, together with tailored loss functions, have been used to improve fault prediction performance, with reported gains in precision, recall, and G-mean compared to baseline methods (Dangut et al., 2023). At the system level, hybrid frameworks integrating sensor data and maintenance logs have demonstrated fault detection rates up to 95% in operational settings (Kabashkin et al., 2025).

Simulation-based and digital twin approaches provide controlled environments for model development and evaluation. Studies using simulated actuator and fuel system data report classification accuracies between approximately 90% and 96%, depending on model type and experimental setup (Kraemer & Villani, 2024; M. Li et al., 2025). These approaches often combine multiple model classes, including neural networks and classical machine learning algorithms, within structured diagnostic frameworks.

To further clarify modality-specific behavior, Table 1 summarizes the comparative performance characteristics observed across the major input modalities. It aggregates representative accuracy ranges, typical datasets, commonly used model classes, and recurring limitations for time-series, image-based, log-based, and simulation-driven approaches.

This comparison complements the detailed discussion in Section 3.1 by providing a modality-level perspective, illustrating that performance differences are not solely attributable to model architecture but are strongly influenced by the nature, quality, and structure of the underlying data.

Table 1: Comparative Performance Across Modalities

Modality	Average Accuracy Range (Representative)	Typical Dataset Type	Common Models Used	Observed Limitations	Citations
Time-series signals (engines, actuators, control surfaces, sensors)	≈90–99% on simulation/offline; ≈90–98% on real/pseudo-real-time (context-dependent)	NASA CMAPSS; actuator rigs; AirLab UAV dataset; multi-flight sensor datasets	CNN, 1D-CNN–BiLSTM, LSTM/AE, ensembles, k-NN/DT/ANN	Domain shift between simulated and real; limited multi-condition coverage; interpretability gaps in deep models	(Kraemer & Villani, 2024; Wu et al., 2024; Young et al., 2025)

Modality	Average Accuracy Range (Representative)	Typical Dataset Type	Common Models Used	Observed Limitations	Citations
Image-based sensor representations	≈98–99% across multiple aircraft/conditions	Real multi-aircraft plus small-scale flight tests	Deep CNNs (VGG16-based, pruned), Grad-CAM/SHAP XAI	Sensitivity to preprocessing and image construction; focus on specific sensors	(J. Jia & Li, 2023; Z. Li et al., 2023)
Log / maintenance / multi-airline operational data	Fault detection rate around 95% (hybrid ATSaaS)	Real maintenance logs, operational metrics from 86 airlines	Rule-based logic, RF, LSTM, hybrid ensembles	Class imbalance, heterogeneous coding; emphasis on system KPIs rather than per-fault metrics	(Dangut et al., 2023; Kabashkin et al., 2025)
Simulation / digital-twin-based datasets	High classification accuracy (often >95%, context-specific)	CMAPSS, fuel-system and actuator digital twins	CNN/LSTM hybrids, ANN, SVM, DT, hybrid model-/data-based	Domain shift, limited operational variability, and reduced real-world representativeness.	(Huang et al., 2023; J. Li et al., 2025; Wu et al., 2024)

To provide a system-level perspective on how different data modalities are processed within AI-based fault detection frameworks, Figure 2 presents a modality-aware pipeline architecture. The figure illustrates how time-series signals, image-based data, maintenance logs, and simulation outputs each require distinct preprocessing, feature extraction, and modeling stages before converging into a unified decision layer. This representation emphasizes that fault detection performance and deployability are strongly dependent on modality-specific design choices, rather than on model architecture alone.

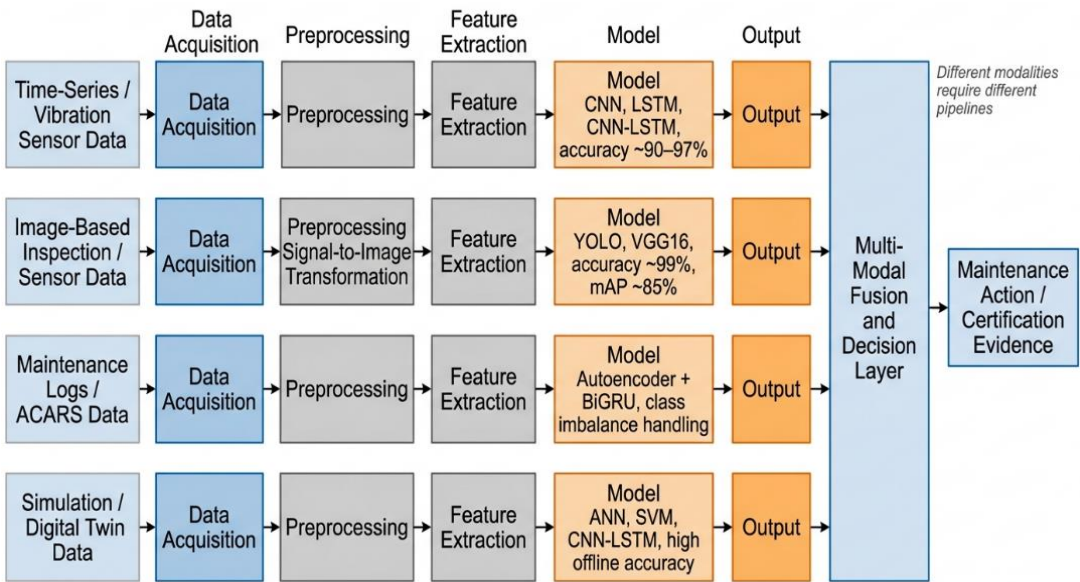


Figure 2: Conceptual Architecture AI Fault Detection Pipelines in Aircraft Systems

Data Provenance: Simulation, Real Flight, and Hybrid Sources

The surveyed literature employs three primary categories of data sources: simulation-based datasets, real-flight or fleet data, and hybrid combinations of both.

Simulation-based studies are widely used for benchmarking and controlled experimentation. For example, models trained on datasets such as CMAPSS or synthetic fuel system simulations report classification accuracies up to approximately 90%, depending on the complexity of the fault scenarios and model architecture (Z. Li et al., 2025; Wu et al., 2024). These datasets provide labeled fault conditions and systematic coverage of operating regimes (R. Donatus et al., 2025).

Real-flight and fleet-level data are used in a smaller but significant subset of studies. Applications include commercial aircraft maintenance logs, air data sensor measurements, and onboard monitoring systems. LSTM-based and CNN-based models applied to such datasets report performance levels around 90% accuracy across multiple aircraft and operating conditions, with additional qualitative evidence of operational effectiveness in maintenance scenarios (Y. Zhao et al., 2022). Visual inspection datasets derived from real aircraft imagery have also been used to train detection models with performance metrics such as mAP around 85% (Connolly et al., 2024).

Hybrid data strategies combine simulated and real datasets to improve coverage and robustness. In actuator diagnostics, frameworks integrating real flight data with model-generated samples report balanced accuracies above 90% (Kraemer & Villani, 2024). Similarly, AI-based health monitoring systems incorporating both simulated and operational data achieve classification accuracies around 97% and demonstrate capability for remaining useful life estimation (Kabashkin, 2025). For maintenance logs, models trained exclusively on real data with imbalanced distributions report improved precision and recall without reliance on synthetic oversampling techniques (Dangut et al., 2023).

To provide a structured view of how data provenance and validation practices are operationalized across the literature, Table 2 summarizes the key datasets used in AI-based aircraft fault detection studies. The table highlights the type of data (real, simulated, or hybrid), application domain, and validation strategies employed, alongside their associated limitations.

This synthesis makes explicit that dataset characteristics and validation design are tightly coupled, and that many reported performance levels are contingent on constrained data regimes rather than fully representative operational environments.

Table 2: Datasets and Validation Strategies in AI-Based Fault Detection

Dataset / Data Source	Type (Real / Simulated / Hybrid)	Application Domain	Validation Approach	Limitations	Citations
Rotor blade vibration dataset (4,921 samples)	Real	Predictive maintenance of rotor systems	Train/test split (3,601/1,320); comparison of multiple models	Single rotor application; limited fleet diversity and long-term drift analysis	(Young et al., 2025)

Dataset / Data Source	Type (Real / Simulated / Hybrid)	Application Domain	Validation Approach	Limitations	Citations
NASA CMAPSS	Simulated	Aircraft engine degradation and fault diagnosis	Offline experiments; comparison across models on single operating condition	Single-condition focus; no direct real-flight validation; potential domain shift	(Huang et al., 2023; Wu et al., 2024)
Multi-aircraft inertial/sensor "state image" database	Real	Air-data sensor fault detection	Cross-condition evaluation on 6 aircraft/flight conditions	Proprietary dataset; limited public reproducibility; confined to air-data sensors	(Y. Zhao et al., 2022)
Multi-aircraft sensor and image dataset (4 aircraft, 5 conditions + model airplane)	Real	Aerospace sensor FDC (image-based)	Cross-aircraft/condition offline validation; additional deployment on model airplane	Focuses on selected sensor suite; long-term operational robustness not demonstrated	(Z. Li et al., 2023)
Electrohydraulic actuator data (simulation + flight records)	Hybrid	Flight-control actuator fault diagnosis	Multi-phase framework: training, operation, improvement; balanced accuracy reported	Performance drop in improvement phase suggests domain shift; still offline	(Kraemer & Villani, 2024)
Simplified fuel-system simulation dataset	Simulated	Multi-fault diagnosis in aircraft fuel system	Supervised offline validation across multiple fault combinations	No real-flight validation; limited operating scenarios	(J. Li et al., 2025)
AirLab failure and anomaly dataset	Real	Autonomous fixed-wing UAV control surface faults	Train/validation/test; evaluation for prediction, detection, classification; real/pseudo-real-time	Restricted to one UAV platform; certification concerns not explored	(Ahmad et al., 2022)
Multi-airline ATSaaS dataset (sensor, maintenance, operations)	Real	Fleet-level maintenance, fault detection, optimization	10-fold cross-validation; comparison across customization, personalization, hybrid	Heterogeneous airline practices; model-level detail abstracted; potential label noise	(Kabashkin et al., 2025)

To systematically conceptualize how AI-based fault detection systems evolve toward operational deployment, Figure 3 presents a validation maturity flowchart. The figure organizes existing approaches along a progressive ladder, from simulation-based concept validation to in-service fleet deployment, highlighting the increasing requirements for robustness, generalization, and certification readiness. This framing emphasizes that most current studies remain concentrated in early stages, while aviation-grade deployment requires progression toward cross-condition, digital twin-integrated, and real-flight validation.

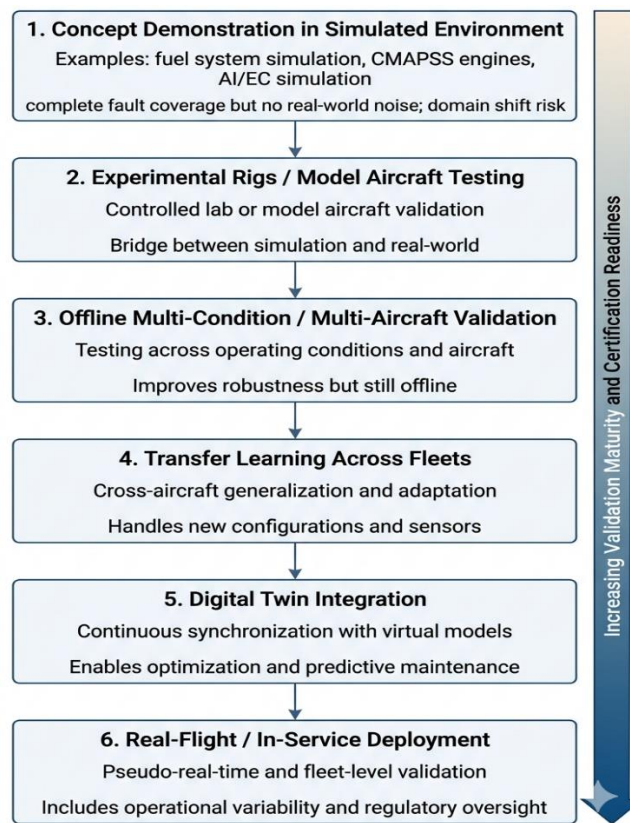


Figure 3: Validation Maturity Flowchart for AI- Based Aircraft Fault Detection

Validation Maturity: Offline, Cross-Condition, and Real-Time Evaluation

Validation approaches in the literature range from offline experiments to cross-condition and near real-time evaluations.

Many studies rely on offline validation using fixed train–test splits or cross-validation on a single dataset. Examples include CNN-based rotor vibration analysis and CNN–LSTM-based air data sensor diagnostics, which report accuracies in the range of 90% to 97% under controlled conditions (Young et al., 2025; Y. Zhao et al., 2022). Image-based diagnostic models using transformed sensor data similarly report high accuracy across predefined operating scenarios (Z. Jia et al., 2023; J. Li et al., 2023).

Cross-condition validation is implemented in studies that evaluate models across multiple aircraft, operating regimes, or environmental conditions. Multi-aircraft datasets and multi-condition experiments demonstrate that high classification accuracy can be maintained across varying scenarios, with reported values exceeding 98% in some cases (J. Li et al., 2023). Evaluations involving multiple aircraft configurations and operational cases provide additional evidence of model generalization (Y. Zhao et al., 2022).

A smaller subset of studies considers real-time or pseudo real-time validation. In autonomous flight applications, LSTM-based models have achieved approximately 93% prediction accuracy with measurable lead times before fault occurrence (Ahmad et al., 2022). Health monitoring frameworks designed for continuous operation report

classification accuracies near 97% under simulated real-time conditions (Kabashkin, 2025). In manufacturing and inspection contexts, real-time deployment of CNN-based models has been evaluated using operational metrics such as defect detection accuracy and reductions in rework time (Shafi et al., 2023).

Structured benchmarking and large-scale cross-validation are less common but present in some studies. For example, fault detection systems evaluated using multi-airline datasets and k-fold cross-validation report detection rates up to 95% (Kabashkin et al., 2025). Hybrid diagnostic frameworks also include comparative validation against alternative model classes using consistent evaluation protocols (Kraemer & Villani, 2024).

Model Class: Traditional, Deep Learning, Hybrid, and Advanced Methods

The literature includes a broad range of model classes, from traditional machine learning to advanced deep learning and hybrid architectures.

Traditional machine learning methods, including support vector machines, decision trees, and logistic regression, are applied primarily in structured diagnostic problems. In aircraft fuel system diagnostics, artificial neural networks achieve accuracies around 90%, while decision trees provide interpretable results with quantified consistency relative to system behavior (Chen et al., 2024; J. Li et al., 2025). Ensemble approaches combining multiple classical models have also demonstrated balanced accuracies up to 96% in actuator diagnostics (Kraemer & Villani, 2024).

Deep learning models dominate applications involving high-dimensional and temporal data. CNNs, LSTMs, and auto encoder-based architectures are widely used for sensor data analysis and fault prediction. Reported performance includes accuracies between 90% and 97% for CNN-based models and up to 100% classification accuracy in specific LSTM-based fault detection tasks (Ahmad et al., 2022; Young et al., 2025; Y. Zhao et al., 2022). Hybrid deep architectures combining convolutional, recurrent, and attention mechanisms further extend these approaches, achieving improved classification performance on benchmark datasets (Wu et al., 2024).

Hybrid models integrate multiple learning paradigms or combine data-driven and rule-based components. Examples include recurrent neural networks combined with self-organizing maps for multi-mode fault detection (Shen & Khorasani, 2020), and auto encoder–recurrent architectures with customized loss functions for rare-event prediction (Dangut et al., 2023). System-level implementations integrate rule-based reasoning with machine learning models, achieving fault detection rates up to 95% in operational platforms (Kabashkin et al., 2025).

Explainable and deployable deep learning models are also reported, particularly in image-based diagnostics. Transfer learning combined with pruning techniques enables lightweight models with high accuracy, supported by visualization methods such as Grad-CAM and SHAP for interpretability (J. Li et al., 2023). Similar approaches are used in sensor fault detection to provide interpretable feature representations and model outputs (Y. Zhao et al., 2022).

Generative and data augmentation approaches appear primarily in the context of autoencoders and simulation-based data generation. These methods support feature learning and class balancing but are not yet widely represented as standalone generative modeling frameworks in the surveyed literature (Kabashkin, 2025).

To consolidate the multi-dimensional analysis presented in this section, Table 3 provides a structured taxonomy of representative AI-based aircraft fault detection studies. The table organizes the literature along the key axes introduced in this survey input modality, data provenance, model class, and validation level—while also summarizing performance metrics, strengths, and limitations for each approach.

Rather than reiterating individual study results, the table enables a compact comparison across heterogeneous systems and datasets, making visible patterns that are not immediately apparent in narrative form. In particular, it highlights how similar performance levels are achieved under different data conditions and validation assumptions, thereby supporting a more nuanced interpretation of reported accuracies.

Table 3: Taxonomy of AI-Based Aircraft Fault Detection

Input Modality	Data Source (Simulation / Real / Hybrid)	Model Class	Validation Level*	Performance Metrics	Key Strength	Key Limitation	Study (Author, Year)
Time-series vibration signals (rotor blades)	Real flight test data	CNN vs other ML/DL	Level 2 — Cross-Condition Validation	Accuracy 97%, Precision 98%, Recall 97%, F1 97%	High accuracy with sensor-rich real data; modality-aligned CNN design	Limited to one application and sensor set; no in-service fleet validation	(Young et al., 2025)
Engine monitoring time-series (raw 1D signals)	Simulation (NASA CMAPSS)	1D-CNN–BiLSTM with CBAM	Level 1 — Offline Single-Scenario Validation	Higher accuracy than baseline models (exact values dataset-specific)	End-to-end learning on raw data; strong benchmark performance	Single operating condition; simulation-only validation; no interpretability yet	(Wu et al., 2024)
Image-based representation of sensor data	Real flight from four aircraft; additional model-airplane flight	Deep CNN (VGG16-based, pruned) with XAI	Level 2 — Cross-Condition Validation	98.99% accuracy (4 aircraft, 5 conditions), 98.44% on model airplane	high accuracy plus Grad-CAM/SHAP-based explainability; lightweight deployment	Focus on specific sensor configuration; limited discussion of long-term in-service drift	(Z. Li et al., 2023)
Time-series “state images” from inertial/sensor data	Real flight data for multiple aircraft/conditions	CNN+LSTM DNN with visualization	Level 2 — Cross-Condition Validation	≈90% accuracy over 24 diverse cases	Robust formulation with scalable performance and CAM-based interpretability	Limited to air-data sensors; no explicit fleet-scale maintenance integration	(Y. Zhao et al., 2022)
Actuator sensor time-series	Hybrid (simulated + real flight records)	Hybrid supervised/unsupervised	Level 2 — Cross-Condition Validation	Balanced accuracy 96% (operation),	Lifecycle-oriented framework; hybrid data use	Still offline; improvement phase	(Kraemer & Villani, 2024)

Input Modality	Data Source (Simulation / Real / Hybrid)	Model Class	Validation Level*	Performance Metrics	Key Strength	Key Limitation	Study (Author, Year)
		(clustering + ensemble ML)		90.4% (improvement)	improves generalization	performance degradation indicates domain shift	
Flight control surface sensor time-series	Real flight (AirLab dataset)	LSTM-based DL	Level 4 — In-Service or Pseudo Real-Time Validation	Failure prediction 93% accuracy; 100% detection and classification	Real-time/pseudo-real-time capability; early failure prediction (~19 s)	Focused on one dataset; certification and explainability not addressed	(Ahmad et al., 2022)
Simulated fuel-system signals	Simulation (B777 fuel model)	ML (ANN, SVM, DT, LR) with XAI	Level 2 — Cross-Condition Validation	ANN accuracy 90%; DT ~lower but highly interpretable	Explicit accuracy–interpretability comparison using LIME/SHAP; domain-grounded	Simulated data only; no real-flight fuel-system validation	(J. Li et al., 2025)
Fuel-system signals (experimental & simulated, review scope)	Simulation and experiment (surveyed)	ML and DL (reviewed)	Not applicable (review)	Not applicable	Identifies need for interpretable, certifiable fuel-system diagnostics	No unified quantitative benchmarking; limited operational evidence	(J. Li et al., 2023)
Mixed: real-time sensor, maintenance logs, operational metrics	Real multi-airline maintenance/operation data	Hybrid: rule-based + ML (RF, LSTM)	Level 2 — Cross-Condition Validation	Fault detection rate 95%; improved downtime and cost metrics	System-level ATSaaS architecture; demonstrates gains over baselines	Platform-level; individual model details and per-fault metrics abstracted	(Kabashkin et al., 2025)

These modality-specific analyses motivate a cross-modal comparison of performance, limitations, and trade-offs, presented in the next section.

Cross-Modal Analysis and Discussion

Alignment with Existing Knowledge and Prior Reviews

The findings from Section 3 are broadly consistent with prior surveys on aircraft health monitoring and predictive maintenance, while extending them through a structured consideration of modality and validation. Existing reviews emphasize the strong diagnostic capability of AI methods alongside persistent concerns regarding explainability, data realism, and system integration (Kabashkin et al., 2025; J. Li et al., 2023). The present synthesis confirms that high classification accuracies, typically in the range of 90% to 99%, are frequently reported across sensor-based and image-based applications, particularly

under controlled experimental conditions (Wu et al., 2024; Young et al., 2025; Y. Zhao et al., 2022).

A key refinement introduced in this survey is the explicit structuring of results along input modality, data provenance, validation maturity, and model class. This organization reveals that performance outcomes are closely linked to the nature of the data rather than solely to the choice of model architecture. For instance, high accuracies reported for CNN and CNN–LSTM models are typically associated with dense, high-quality time-series sensor data, while image-based transformations combined with transfer learning similarly benefit from well-defined preprocessing pipelines (J. Li et al., 2023; Y. Zhao et al., 2022). In contrast, studies based on maintenance logs and fleet-level data exhibit more modest and variable performance due to data sparsity and imbalance (Dangut et al., 2023; Kabashkin et al., 2025).

Another point of alignment with recent literature is the growing recognition of validation as a critical factor. While earlier surveys often emphasize model performance, the present analysis shows that only a limited number of studies incorporate cross-condition, multi-aircraft, or near real-time validation (Ahmad et al., 2022; Kraemer & Villani, 2024; Y. Zhao et al., 2022). This observation is consistent with maintenance system research that highlights robustness across operational scenarios as a prerequisite for deployment.

Finally, the integration of explainable AI approaches and hybrid system architectures reflects an emerging trend in the literature (R. E. Donatus & Uhiah, 2026). Studies combining interpretable models, visualization techniques, and rule-based components demonstrate how performance and traceability can be jointly addressed (Kabashkin et al., 2025; J. Li et al., 2025). This supports the broader conclusion that effective fault detection systems increasingly balance predictive capability with transparency and validation rigor.

Cross-Modal Insights and Contradictions

The comparative analysis highlights clear differences in performance and characteristics across data modalities. Time-series sensor data consistently support high diagnostic accuracy, particularly in applications involving engines, actuators, and air data systems. Models operating on such data benefit from well-defined temporal patterns and structured measurements, enabling stable classification performance across multiple studies (Ahmad et al., 2022; Young et al., 2025; Y. Zhao et al., 2022).

Image-based approaches provide an alternative pathway to high performance by transforming signals into spatial representations or by directly analyzing visual inspection data (R. E. Donatus & Uhiah, 2026). Methods based on time–frequency images and transfer learning architectures achieve performance comparable to, and in some cases exceeding, time-series approaches when evaluated under similar conditions (Z. Jia et al., 2023; J. Li et al., 2023). Visual inspection systems, while addressing different diagnostic tasks, demonstrate effective detection of structural defects using object detection models (Connolly et al., 2024; R. E. Donatus et al., 2025).

Log and maintenance data present a distinct set of challenges. These datasets are typically sparse, highly imbalanced, and less directly linked to physical system behavior. Although advanced architectures improve detection metrics, overall performance remains more variable compared to sensor-based methods (Dangut et al., 2023; Kabashkin et al., 2025). Simulation and digital twin data enable systematic experimentation and coverage of fault scenarios, but introduce uncertainty regarding transfer to real-world operations. Studies based on simulated datasets report strong performance, yet often acknowledge the need for further validation under operational conditions (J. Li et al., 2025; Wu et al., 2024). Hybrid approaches combining simulated and real data attempt to address this limitation, showing improved robustness in several applications (Kabashkin et al., 2025; Kraemer & Villani, 2024).

Taken together, these observations indicate that different modalities contribute complementary strengths. High accuracy is commonly achieved in controlled sensor and image-based settings, whereas log-based and hybrid data approaches provide closer alignment with operational environments but exhibit greater variability in performance. The disparity between simulation-based performance and real-world operational outcomes is summarized in Figure 4. While many studies report accuracies exceeding 90–99% under controlled or benchmark conditions, performance in real-flight and fleet-level environments is influenced by noise, domain shift, data imbalance, and evolving operational conditions. The figure highlights the key sources of this degradation and underscores the necessity of hybrid datasets, transfer learning, and real-flight validation for certification-relevant fault detection systems.

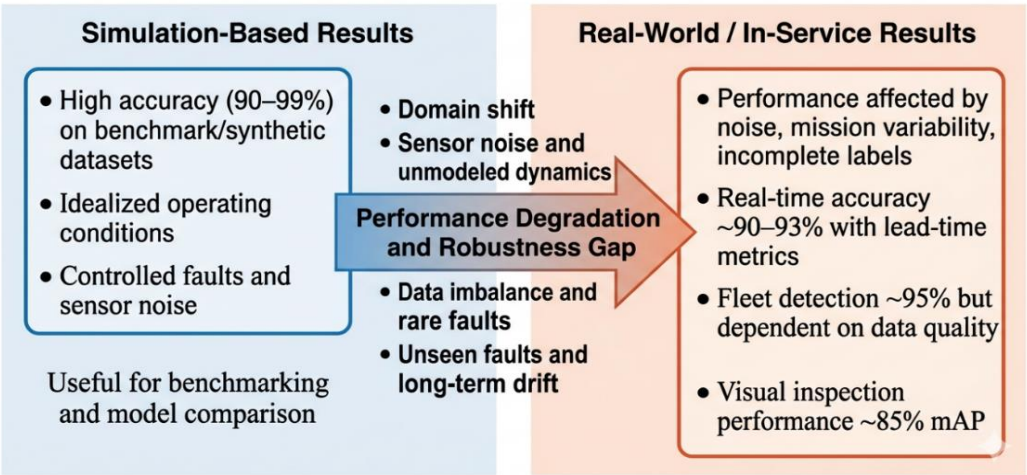


Figure 4: Simulation VS Real- World Performance Gap in Aircraft Fault Detection
As illustrated in Figure 4, the degradation from simulation to operational environments is not merely statistical but structurally driven by data and system-level constraints

Trade-Off Analysis

Three major trade-offs emerge from the surveyed literature: accuracy versus interpretability, performance versus computational cost, and simulation-based performance versus operational robustness.

The fundamental trade-offs governing AI-based fault detection are summarized in Figure 5, which illustrates the competing objectives of predictive accuracy, interpretability, and computational efficiency. The positioning of different model classes within this space highlights that no single approach simultaneously optimizes all three dimensions, reinforcing the need for balanced, application-driven model selection in safety-critical aviation contexts.

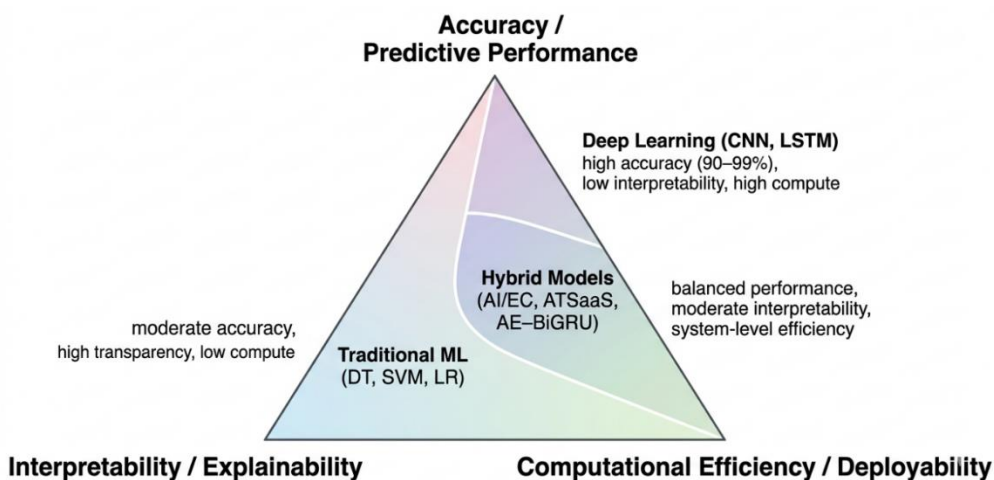


Figure 5: Trade-Off Triangle for AI Model in Aircraft Fault Detection

The balance between accuracy and interpretability remains a central consideration. Deep learning models, including CNNs and LSTMs, consistently deliver high predictive performance on complex datasets (Ahmad et al., 2022; R. Donatus et al., 2025; Wu et al., 2024). However, traditional machine learning methods and rule-based systems provide clearer decision structures and are often preferred in contexts requiring transparency and traceability. For example, decision tree models demonstrate strong alignment with engineering reasoning despite lower classification accuracy compared to neural networks (Ceken & Tuncal, 2026; Kefalas et al., 2022; J. Li et al., 2025). Explainable deep learning techniques, including visualization and feature attribution methods, provide a partial bridge between these approaches (J. Li et al., 2023; Y. Zhao et al., 2022).

Computational requirements represent a second important trade-off. High-capacity deep learning models require substantial resources for training and inference, which can limit their suitability for onboard or real-time applications. Techniques such as model pruning and architecture optimization reduce computational complexity while maintaining strong predictive performance (J. Li et al., 2023). In contrast, simpler models and hybrid systems may offer reduced computational burden with acceptable performance in operational contexts (Kabashkin, 2025; Shafi et al., 2023).

A third trade-off arises between performance achieved on simulated datasets and robustness in real-world deployment. Simulation-based studies provide controlled environments for model development and benchmarking, often resulting in high reported accuracies (Wu et al., 2024). However, real-flight and fleet-level applications introduce additional variability, including noise, missing data, and evolving operating conditions. Hybrid approaches that incorporate both simulated and real data represent an intermediate strategy, improving robustness while retaining coverage of rare fault scenarios (Kabashkin, 2025; Kraemer & Villani, 2024).

Building on the trade-off analysis discussed above, Table 4 synthesizes the relationships between model classes and key engineering considerations, including accuracy, interpretability, computational cost, data requirements, deployment feasibility, and certification readiness.

The table serves as a practical reference for system designers by mapping algorithmic choices to operational constraints, highlighting that no single model class dominates across all criteria. Instead, it reinforces the need for context-aware selection strategies that balance performance with explainability, resource constraints, and regulatory requirements.

Table 4: Trade-off Analysis of AI Models for Aircraft Fault Detection

Model Type	Accuracy (Typical)	Interpretability	Computational Cost	Data Requirement	Deployment Feasibility	Certification Readiness	Citations
Traditional ML (SVM, RF, DT, LR, k-NN)	Moderate-high (e.g., DT slightly below ANN; RF strong on logs)	High for DT/LR; moderate for SVM/RF with XAI	Generally low-moderate	Performs well on small-medium labelled datasets	Suitable for onboard or ATS platforms with tight resource budgets	Favourable due to transparency and simpler assurance arguments	(Dangut et al., 2023; Kraemer & Villani, 2024; J. Li et al., 2025)
Deep CNN / 1D-CNN	High (up to ~97-99% in several studies)	Low-moderate; improved with CAM/Grad-CAM	Moderate-high, depending on depth	Requires large, well-curated datasets and careful preprocessing	Feasible on ground servers; embedded deployment needs pruning/acceleration	Challenging; needs XAI and robust validation evidence	(Z. Li et al., 2023; Wu et al., 2024; Y. Zhao et al., 2022)
LSTM / Bi-LSTM / AE	High for temporal prediction and classification (~90-99%)	Low; sequence attention sometimes available	High for long sequences and large fleets	Long time-series with diverse conditions	Strong for ground-based fleet monitoring; more constrained onboard	Requires work on explainability and handling of non-stationarity	(Ahmad et al., 2022; Wu et al., 2024; Zhang et al., 2025)
Hybrid supervised-unsupervised (e.g., clustering + ensemble)	High balanced accuracy (~90-96%)	Moderate; components interpretable, pipeline more complex	Moderate	Needs both labelled and unlabelled data, including simulated cases	Attractive for lifecycle-oriented platforms (operation + improvement phases)	Promising because physical reasoning can be embedded, but complexity adds assurance burden	(Kabashkin et al., 2025; Kraemer & Villani, 2024; Shen & Khorasani, 2020)

Model Type	Accuracy (Typical)	Interpretability	Computational Cost	Data Requirement	Deployment Feasibility	Certification Readiness	Citations
Physics-informed / digital-twin-based	High in controlled scenarios (values dataset-specific)	Higher than pure DL; grounded in physical variables	Moderate–high (simulation + learning)	Requires validated models and calibration data	Well suited for engineering analysis and what-if studies; online use depends on runtime	Alignment with physics improves trust; still limited operational evidence	(Brunton et al., 2021; Ezhilarasu et al., 2021; Huang et al., 2023)
XAI-enhanced models (e.g., pruned CNN with Grad-CAM/SHAP; interpretable fuel-system ML)	High (e.g., ~99% for sensor FDC, ~90% for fuel faults)	Explicitly improved via LIME/SHAP/CAM	Slightly higher than base models due to analysis overhead	Same as underlying model + need for explanation data	Suitable for both engineering analysis and decision support; onboard feasible for lightweight variants	Better positioned for certification due to traceable explanations	(J. Li et al., 2023, 2025; Y. Zhao et al., 2022)

Implications for Aviation Safety and Certification

The results indicate that AI-based fault detection has significant potential to enhance aviation safety and maintenance efficiency. High detection accuracy and early fault prediction capabilities support proactive maintenance strategies and improved system reliability (Ahmad et al., 2022). System-level implementations integrating AI into maintenance workflows demonstrate measurable improvements in fault detection rates and operational efficiency (Kabashkin, 2025; Kabashkin et al., 2025).

At the same time, the requirements for certification introduce additional constraints. Regulatory acceptance depends on the availability of reliable evidence demonstrating consistent performance under operational conditions, as well as clear and traceable decision processes (R. E. Donatus & Uhiah, 2026). While some studies incorporate explainability techniques and structured validation, many focus primarily on performance metrics without addressing these requirements in detail (J. Li et al., 2023; Ofoegbu & Josiah, 2026; Y.-P. Zhao & Chen, 2022).

Validation practices also vary widely. Although some works include multi-condition testing, cross-validation, and pseudo real-time evaluation, these approaches are not yet standard across the field (Ahmad et al., 2022; Kabashkin et al., 2025). Evaluation under out-of-distribution conditions remains limited, despite its importance for safety-critical systems. Overall, the literature demonstrates strong technical capability but only partial alignment with certification requirements. Bridging this gap will require greater emphasis on explainability, robust validation, and integration with established safety assurance processes.

Key Research Gaps and Unresolved Challenges

Several important gaps emerge from the analysis:

Data availability remains a primary challenge, particularly for rare fault conditions. While simulation and digital twin approaches provide extensive datasets, real-world labeled data are limited and often difficult to access. Studies using maintenance logs highlight the

impact of class imbalance and data heterogeneity on model performance (Dangut et al., 2023).

Validation practices also require further development. Many studies rely on static offline evaluations, with limited consideration of long-term operational variability or cross-fleet generalization (J. Jia & Li, 2023; Wu et al., 2024; Young et al., 2025). Although some works incorporate more advanced validation strategies, systematic and standardized protocols are not yet widely adopted (Kabashkin et al., 2025; Y.-P. Zhao & Chen, 2022).

Deployment considerations are addressed in relatively few studies. Issues such as computational constraints, real-time processing, and system integration are critical for practical implementation but are often treated as secondary concerns (Kabashkin, 2025; J. Li et al., 2023; Shafi et al., 2023). In addition, the interaction between human operators and AI-based diagnostic systems remains underexplored.

Finally, there is a lack of standardized benchmarks that reflect real operational conditions across multiple modalities. Existing datasets support comparative evaluation within specific domains but do not capture the full complexity of aircraft systems, including heterogeneous data sources and evolving operational environments (Wu et al., 2024).

Addressing these challenges will be essential for advancing AI-based fault detection from research prototypes to operational and certifiable systems.

To consolidate the cross-cutting challenges identified in this survey, Table 5 presents a structured synthesis of key research gaps across data, model design, deployment, and certification dimensions. Each category is linked to supporting evidence from the literature, its impact on aviation systems, and corresponding research directions.

This structured mapping moves beyond descriptive analysis by explicitly connecting observed limitations to their operational implications and to targeted avenues for future work.

Table 5: Synthesis of Research Gaps and Open Challenges

Category	Observed Issue	Evidence from Literature	Impact on Aviation Systems	Suggested Research Direction	Citations
Data	Overreliance on simulated or single-condition datasets	Digital-twin and CMAPSS studies emphasize simulation; many works consider only one operating regime	Models risk brittleness under real operational variability; safety margins hard to quantify	Develop standardized multi-condition, multi-fleet benchmarks; integrate real-flight data into digital-twin workflows	(Huang et al., 2023; M. Li et al., 2025; Z. Li et al., 2023; Wu et al., 2024)
Data	Data imbalance and sparse labels for rare faults	Fleet-log and maintenance-platform studies highlight rare failures and heterogeneous records	Elevated false-negative or false-positive rates; reduced trust in alerts	Systematic use of imbalance handling, synthetic/augmented data, and semi-supervised methods tailored to aviation	(Dangut et al., 2023; Kabashkin et al., 2025)

Category	Observed Issue	Evidence from Literature	Impact on Aviation Systems	Suggested Research Direction	Citations
Model	Accuracy–interpretability trade-off in safety-critical decisions	Fuel-system case shows ANN vs DT differences; DNN-based FDC needs XAI; reviews stress explainability	High-accuracy black-box models are difficult to justify to certifiers and maintainers	Co-design of accuracy and interpretability; physics-informed and XAI-enhanced models with quantified explanation quality	(Brunton et al., 2021; J. Li et al., 2025; Z. Li et al., 2023; Y. Zhao et al., 2022; Zolghadri, 2024)
Model	Limited evaluation under domain shift and transfer scenarios	Many studies limited to single platforms; transfer learning only occasionally considered	Poor portability across aircraft types and conditions; costly re-validation	Structured transfer-learning and domain-adaptation studies across platforms and fleets	(Huang et al., 2023; Kraemer & Villani, 2024; Wu et al., 2024)
Deployment	Gap between offline metrics and in-service performance	High offline accuracies contrast with few pseudo-real-time/fleet deployments	Uncertain operational reliability; difficulty estimating safety benefits	More pseudo-real-time pilots, longitudinal field trials, and integration with PHM/ATS platforms	(Ahmad et al., 2022; Kabashkin, 2025; Kabashkin et al., 2025; Zolghadri, 2024)
Deployment	Computational and integration constraints for embedded/onboard use	High-capacity CNN/LSTM models; real-time manufacturing inspection and health monitoring frameworks	Onboard avionics may not support large models; latency constraints	Model compression, edge-optimized architectures, and partitioned on-aircraft/ground processing designs	(Kabashkin et al., 2025; J. Li et al., 2023; Shafi et al., 2023; Zhang et al., 2025)
Certification	Lack of consolidated criteria for AI assurance and evidence	Certification-oriented works highlight missing guidance; reviews stress need for robust validation	Slows adoption of AI-based FD; risk of inconsistent approval processes	Joint development of AI-specific assurance cases, validation maturity scales, and explainability requirements	(Brunton et al., 2021; Luettig et al., 2024; Sridhar et al., 2025)
Certification	Limited modality-aware linkage between fault models and safety arguments	Reviews and case studies often omit modality/provenance in safety discussion	Safety cases overlook how modality and data source affect uncertainty	Modality-aware certification frameworks connecting data provenance, validation maturity, and failure-mode coverage	(Kwakye et al., 2024; J. Li et al., 2023; Zolghadri, 2024)

The identified trade-offs are translated into deployment and certification implications in the following section

To complement the qualitative comparisons, Figure 6 provides a consolidated view of reported performance across input modalities. Rather than presenting strict meta-analytic estimates, the figure summarizes indicative performance ranges derived from representative studies, highlighting those higher reported accuracies are often associated with structured or simulated data under controlled validation settings. In contrast,

modalities closer to operational reality such as maintenance logs and visual inspection exhibit more moderate but practically meaningful performance due to noise, imbalance, and environmental variability.

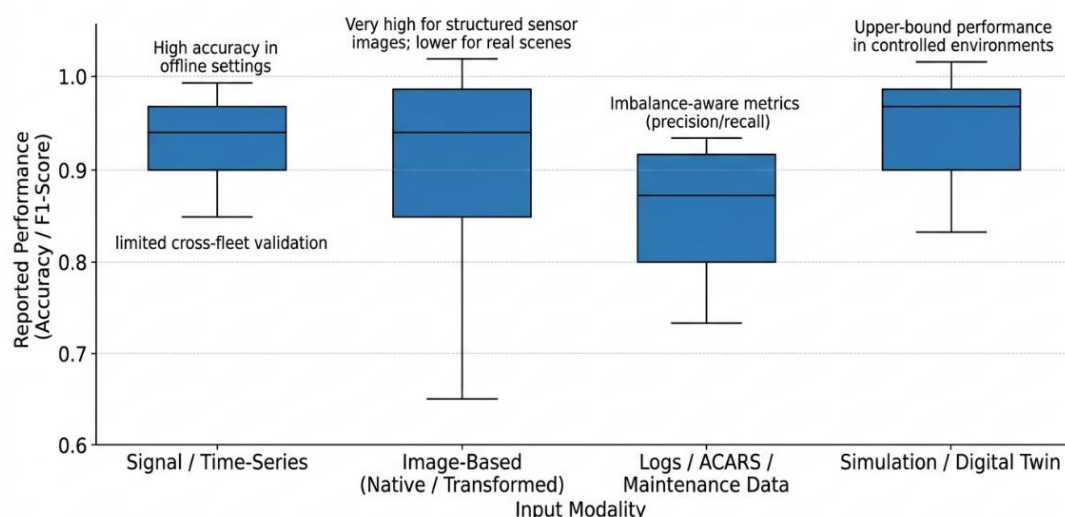


Figure 6: Indicative Performance Ranges Across Input Modalities Based on Representative Studies

The ranges shown are illustrative and derived from representative studies reviewed in this survey rather than from a formal statistical meta-analysis. Reported performance varies substantially with data provenance and evaluation maturity. As shown in Figure 6, time-series and image-based modalities generally occupy the upper performance range, whereas log-based and operational datasets exhibit lower but more operationally realistic metrics, reinforcing the need to interpret reported accuracy within the context of dataset origin and evaluation conditions. Performance also depends strongly on validation maturity, as higher reported scores frequently correspond to controlled, offline, or simulation-based evaluation settings rather than real operational environments.

Engineering Implications and Deployment Framework

Modality-Driven Model Selection

The synthesis demonstrates that model selection in aircraft fault detection must be governed by data modality, operational context, and certification requirements rather than by peak accuracy alone. For high-resolution time-series sensor data, such as those from engines, actuators, and air data systems, convolutional and recurrent architectures consistently provide strong performance. CNNs applied to rotor vibration signals achieve approximately 97% accuracy with high F1-scores, while LSTM-based frameworks enable both accurate fault classification and early prediction with meaningful lead times in flight and fleet applications (Du et al., 2024). Hybrid CNN-LSTM architectures further

demonstrate robustness across multiple aircraft and operating conditions (Y. Zhao et al., 2022). These models are therefore well suited for onboard or near real-time applications where temporal dependencies are critical.

In cases where signals can be transformed into structured image representations, such as time–frequency maps or residual images, two-dimensional convolutional networks and transfer learning approaches become effective. Multi-channel image fusion techniques improve diagnostic performance by integrating complementary signal domains (Z. Jia et al., 2023), while pruned transfer learning models achieve strong predictive performance with reduced computational complexity and successful transfer across aircraft configurations (Z. Li et al., 2023). These approaches are particularly suitable for edge or centralized processing environments where preprocessing overhead is manageable.

For log and maintenance message data, which are typically sparse and highly imbalanced, hybrid temporal architectures are more appropriate. Models that combine autoencoders with recurrent layers and imbalance-aware loss functions have demonstrated improved detection of rare failure events in real airline datasets (Dangut et al., 2023). At the system level, integrated architectures that combine rule-based reasoning with machine learning and sequence models achieve superior fault detection and predictive maintenance performance compared to single-method approaches (Kabashkin et al., 2025).

Simulation and digital twin environments support both traditional and deep learning approaches, offering controlled conditions for model development. However, such models should be treated as preliminary unless validated against real operational data. In safety-critical contexts, interpretable models remain essential. For example, decision trees may offer slightly lower accuracy than neural networks but provide clearer alignment with engineering reasoning, making them more suitable for certification-oriented applications (J. Li et al., 2025).

Data Requirements and Preprocessing Considerations

Effective deployment depends on realistic assumptions about data availability and quality. Time-series models require sufficiently long and well-sampled sequences that capture both nominal and fault evolution. Approaches such as LSTM autoencoders rely on stable baseline behavior to construct meaningful health indices (Ning et al., 2021), while high-dimensional vibration-based models depend on careful signal alignment, normalization, and sensor selection to achieve optimal performance.

Image-based pipelines introduce additional preprocessing requirements. Signal-to-image transformations must be consistent and reproducible, particularly when using time–frequency representations or stacked sensor images (Z. Jia et al., 2023). Data augmentation and model compression techniques, such as channel pruning, are necessary to ensure scalability and deployment feasibility without sacrificing diagnostic performance (Z. Li et al., 2023).

For fleet-scale maintenance logs, preprocessing becomes more complex due to noise, sparsity, and extreme class imbalance. Studies show that specialized techniques, including

focal loss and representation learning through autoencoders, are required to extract meaningful patterns (Dangut et al., 2023). At the platform level, combining heterogeneous data sources such as real-time sensor streams and historical logs requires robust data integration pipelines and balancing strategies, including synthetic augmentation and digital twin support (Kabashkin, 2025; Kabashkin et al., 2025).

Hybrid datasets combining simulated and real data require structured feature engineering. Feature selection and clustering methods help construct interpretable representations and improve the effectiveness of hybrid learning pipelines (Kraemer & Villani, 2024). Across all modalities, deployment-ready systems must implement standardized preprocessing pipelines that are physically meaningful and reproducible.

Validation Strategies for Safety-Critical Systems

The findings emphasize that validation maturity is a primary determinant of deployment readiness. High accuracy obtained from simple train–test splits is insufficient for safety-critical systems. Instead, validation must be structured across multiple levels.

First, statistically grounded evaluation methods such as k-fold cross-validation provide robust performance estimates within a given dataset. Large-scale implementations across multiple airlines or platforms offer stronger evidence of reliability (Kabashkin et al., 2025).

Second, cross-condition and cross-platform validation is essential. Models must demonstrate consistent performance across different aircraft, operating regimes, and environmental conditions. Studies that evaluate models across multiple aircraft configurations and flight conditions provide more credible evidence for scalability (Kefalas et al., 2022; J. Li et al., 2023; Y. Zhao et al., 2022).

Third, real-time or pseudo real-time validation is required to assess operational usefulness. Metrics such as time to detection, prediction lead time, and response latency must be explicitly quantified. Frameworks that incorporate these elements demonstrate closer alignment with real mission requirements (Kabashkin, 2025).

For certification, validation strategies must align with lifecycle-oriented evaluation practices. This includes testing under operational variability, assessing robustness to unseen conditions, and ensuring traceability of model decisions. Current literature shows progress in this direction, but such practices remain limited to a subset of studies.

Deployment Constraints: Onboard vs. Ground Systems

Deployment decisions are strongly influenced by computational and integration constraints. Onboard systems require lightweight, efficient models capable of real-time inference under strict resource limitations. While high-capacity deep models provide strong performance, they often require compression techniques such as pruning or distillation to meet onboard constraints (J. Li et al., 2023).

For UAVs and smaller platforms, resource limitations are even more restrictive. Carefully designed recurrent models can meet real-time requirements when input dimensionality is

controlled, but more complex architectures may need to be offloaded to ground systems (Bronz et al., 2020).

Ground-based and cloud-supported systems offer greater computational flexibility, enabling the deployment of more complex models and multimodal fusion strategies. Maintenance platforms that integrate fleet-wide data and optimization frameworks are typically implemented in such environments (Kabashkin, 2025; Kabashkin et al., 2025).

A practical deployment strategy emerging from the literature is a hierarchical approach: lightweight, explainable models operate onboard for immediate detection, while more complex analytics are performed on ground systems to support deeper diagnostics and maintenance planning.

The deployment architecture of AI-based fault detection systems is illustrated in Figure 7, which distinguishes between onboard and ground-based processing roles. The figure formalizes a split architecture in which lightweight, latency-sensitive models operate onboard within strict computational and safety constraints, while more complex analytics, retraining, and fleet-level learning are performed on ground or cloud platforms. This separation reflects practical limitations in avionics environments and highlights the importance of feedback loops for continuous model improvement and certification alignment.

Deployment Architecture for Onboard vs Ground AI in Aircraft Fault Detection

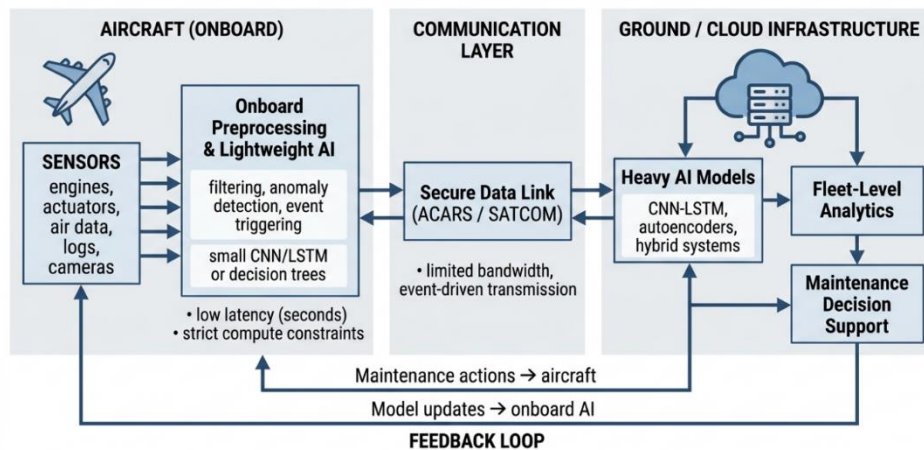


Figure 7: Deployment Architecture for Onboard vs Ground AI in Aircraft Fault Detection

Deployment Readiness Framework

To translate research outcomes into operational systems, a structured deployment readiness framework is required. Based on the survey, five key dimensions define readiness:

- **Modality–Model Fit:** Alignment between data type and model architecture, ensuring that model capabilities match the characteristics of the input modality.

- **Data Provenance and Coverage:** Clear documentation of whether data are simulated, experimental, or operational, and the extent to which they represent real-world conditions.
- **Validation Maturity:** Progression from basic offline evaluation to cross-platform validation and real-time testing with operational metrics.
- **Computational Feasibility:** Assessment of whether models can meet the constraints of onboard or ground-based deployment environments.
- **Interpretability and Certification Alignment:** Availability of transparent decision mechanisms or explainability tools that support engineering validation and regulatory approval.

Each candidate system can be systematically evaluated across these dimensions to determine its readiness level and identify gaps requiring further development. This structured approach enables a systematic transition from high-performing experimental models to certifiable, deployable fault detection systems.

From a certification perspective, AI-based fault detection systems must demonstrate consistent performance across operating conditions, clear traceability between inputs and decisions, and alignment with safety-critical assurance requirements. These considerations directly influence model selection, validation strategy, and deployment architecture.

Future Research Directions

Data and Benchmarking Limitations

The research directions outlined in this section build directly on the structured gaps summarized in Table 5, extending them into actionable development pathways for data, models, validation, and deployment.

The current literature is characterized by fragmented datasets and subsystem-specific studies, many of which are proprietary and not publicly accessible. While simulation environments, component-level rigs, and controlled experiments provide valuable insights, they do not constitute standardized benchmarks that enable consistent comparison across methods. Studies based on fleet-scale maintenance logs further highlight challenges related to class imbalance and heterogeneous data structures (Dangut et al., 2023; Kabashkin et al., 2025).

Future research should focus on the development of structured, modality-aware benchmark suites that integrate simulated and real flight data across representative aircraft subsystems. Such benchmarks should include standardized data splits, clearly defined fault taxonomies, and consistent evaluation protocols aligned with lifecycle considerations. In parallel, the creation of de-identified, open-access fleet maintenance datasets would enable more realistic evaluation of rare fault detection. Performance metrics should also evolve beyond accuracy to include operational indicators such as time to detection, false alarm cost, and maintenance impact, reflecting the system-level evaluation approaches already demonstrated in advanced maintenance platforms.

Simulation–Reality Gap and Domain Adaptation

A persistent limitation in current research is the gap between performance in simulated environments and robustness in real operational settings. High accuracy results obtained from digital twins and benchmark datasets do not necessarily translate to real-world reliability, particularly under varying operating conditions and sensor noise. Although hybrid approaches that combine simulated and real data offer partial mitigation, systematic evaluation of performance degradation during transfer remains limited (Kabashkin, 2025; Kraemer & Villani, 2024).

Future work should prioritize controlled domain shift studies, where models trained on simulated data are rigorously evaluated on real or staged operational data. Transfer learning and domain adaptation strategies should be formalized as standard practice, leveraging simulation for pretraining and real-world data for fine-tuning. In addition, adaptive digital twin frameworks that incorporate streaming operational data to continuously update system models represent a promising direction toward closing the simulation–reality gap.

Explainable and Certifiable AI

Explainability remains a critical requirement for the adoption of AI in safety-critical aerospace systems. While some studies incorporate interpretable models or post hoc explanation techniques, many high-performing deep learning models lack transparent decision mechanisms. This limits their suitability for certification, where traceability and engineering validation are essential.

Future research should emphasize the co-design of models and explanation mechanisms, ensuring that diagnostic outputs are accompanied by physically meaningful interpretations. These explanations should be integrated into higher-level reasoning frameworks to support system-level decision-making and maintenance planning. Furthermore, explainability methods must be aligned with emerging certification paradigms, enabling their incorporation into formal assurance processes and safety cases.

Lightweight and Onboard Efficient Models

Deployment constraints, particularly in onboard and embedded systems, necessitate models that balance performance with computational efficiency. While recent work demonstrates that model compression techniques such as pruning and distillation can significantly reduce complexity without substantial loss of accuracy, systematic design methodologies for lightweight models remain underdeveloped.

Future research should explore hardware-aware model design, including quantization, architecture optimization, and energy-efficient inference tailored to avionics and UAV platforms. Comparative evaluation of latency, energy consumption, and inference performance across onboard, edge, and cloud environments is also required. Additionally, tiered system architectures, where lightweight onboard models are complemented by

more complex ground-based analytics, should be formalized as a standard deployment paradigm.

Multimodal and Digital Twin Integration

Most existing fault detection systems remain focused on single data modalities, despite evidence that multimodal integration can enhance robustness and diagnostic accuracy. Initial efforts combining sensor data, image representations, maintenance logs, and digital twin outputs demonstrate the potential of integrated approaches, particularly for complex and evolving fault scenarios.

Future work should develop standardized multimodal fusion frameworks that explicitly combine heterogeneous data sources within unified architectures. These frameworks should leverage digital twin technologies to incorporate physical knowledge and enable continuous system adaptation. At the fleet level, distributed learning approaches, including federated learning, offer opportunities for collaborative intelligence while preserving data privacy. Integrating multimodal diagnostics into predictive maintenance decision frameworks will further strengthen the link between fault detection and operational outcomes.

Conclusion

This survey has provided a structured, modality-aware synthesis of AI-driven fault detection in aircraft systems, highlighting both the significant progress achieved and the limitations that constrain operational deployment. By organizing the literature along the dimensions of input modality, model class, data provenance, and validation maturity, the study clarifies that many reported high accuracies are achieved under controlled conditions, while performance in operationally realistic environments remains less well established.

The primary contribution lies in translating this structured analysis into an engineering-oriented deployment framework. By explicitly linking model selection, data requirements, validation strategies, and deployment constraints, the work offers a practical pathway from algorithm development to real-world implementation. The inclusion of a deployment readiness framework further enables systematic assessment of candidate solutions in terms of robustness, feasibility, and certification alignment.

Looking forward, advancing AI-based fault detection in aerospace will depend on addressing key challenges related to data availability, validation rigor, explainability, and deployment efficiency. Progress will require coordinated efforts to develop representative benchmarks, bridge the simulation–reality gap, design certifiable and interpretable models, and integrate multimodal data within scalable architectures. With these developments, AI-based fault detection can transition from high-performing experimental systems to reliable, certifiable components of next-generation aircraft health management systems. This progression is essential for bridging the gap between algorithmic accuracy and certified airworthiness in safety-critical aviation systems.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement: No new data were generated or analyzed in this study. All sources referenced are publicly available in the cited literature.

Conflict of Interest: The authors declare no known competing financial interests or personal relationships that could have influenced the work reported in this manuscript.

References

- Ahmad, M. W., Akram, M. U., Ahmad, R., Hameed, K., & Hassan, A. (2022). Intelligent framework for automated failure prediction, detection, and classification of mission critical autonomous flights. *ISA Transactions*, 129, 355–371. <https://doi.org/10.1016/j.isatra.2022.01.014>
- Ahmed, W. (2025). Artificial intelligence in aviation: A review of machine learning and deep learning applications for enhanced safety and security. *Premier Journal of Artificial Intelligence*, 3(1), 100013. <https://doi.org/10.70389/PJAI.100013>
- Azanaw, G. M. (2024). Integrating AI in Structural Health Monitoring (Shm): A Systematic Review on Advances, Challenges, and Future Directions. *I-Manager's Journal on Structural Engineering*, 13(3). <https://doi.org/10.26634/jste.13.3.21791>
- Bronz, M., Baskaya, E., Delahaye, D., & Puechmorel, S. (2020). Real-time fault detection on small fixed-wing UAVs using machine learning. *DASC 2020 AIAA/IEEE 39th Digital Avionics Systems Conference*, ISBN-978. <https://doi.org/10.1109/DASC50938.2020.9256800>
- Brunton, S. L., Kutz, J. N., Manohar, K., Aravkin, A. Y., Morgansen, K., Klemisch, J., Goebel, N., Buttrick, J., Poskin, J., Blom-Schieber, A. W., Hogan, T., & McDonald, D. (2021). Data-driven aerospace engineering: Reframing the industry with machine learning. *AIAA Journal*, 59(8), 2820–2847. <https://doi.org/10.2514/1.J060131>
- Ceken, S., & Tuncal, A. (2026). Artificial intelligence for sustainable aviation: a review on operational implementations and future perspectives. *The Aeronautical Journal*, 1–46. <https://doi.org/10.1017/aer.2026.10138>
- Chen, G., Yuan, J., Zhang, Y., Zhu, H., Huang, R., Wang, F., & Li, W. (2024). Enhancing reliability through interpretability: A comprehensive survey of interpretable intelligent fault diagnosis in rotating machinery. *IEEE Access*, 12, 103348–103379. <https://doi.org/10.1109/ACCESS.2024.3430010>
- Connolly, L., Garland, J., & Tobin, E. (2024). Deep-Learning-Based Defect Detection for Light Aircraft With Unmanned Aircraft Systems. *IEEE Access*, June. <https://doi.org/10.1109/ACCESS.2024.3412204>
- Dangut, M. D., Jennions, I. K., King, S., & Skaf, Z. (2023). A rare failure detection model for aircraft predictive maintenance using a deep hybrid learning approach. *Neural Computing and Applications*, 35(4), 2991–3009. <https://doi.org/10.1007/s00521-022-07167-8>
- Dmitriev, K., Schumann, J., & Holzapfel, F. (2021). Toward certification of machine-learning systems for low criticality airborne applications. *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, 1–7. <https://doi.org/10.1109/DASC52595.2021.9594467>
- Donatus, R. E. (2025). Comprehensive Review of Artificial Intelligence for Image- and Signal-Based Nondestructive Testing in Aerospace Structures. *Journal of Computers, Mechanical and Management*, 4(6), 1–14. <https://doi.org/10.57159/jcmm.4.6.25230.1>
- Donatus, R. E., Ubadike, O. C., Bonet, M. U., & Iyaghiba, S. D. (2025). Automated Aircraft Structural Defect Detection Using Deep Learning and Computer Vision. *Mekatronika Journal of Mechatronics and Intelligent Manufacturing*, 7(2), 108–123. <https://doi.org/10.15282/mekatronika.v7i2.11787>

- Donatus, R. E., & Uhiah, L. O. (2026). Certification And Trustworthy Deployment of Explainable And Physics-Informed Artificial Intelligence For Aerospace Non-Destructive Testing. *International Journal of Engineering Processing & Safety Research*, 11(5), 1–30. <https://doi.org/10.70382/caijepsr.v11i5.010>
- Donatus, R., Uhiah, L., & Olanrewaju, A. (2025). Attention-Enhanced Temporal Modeling for Prognostics of Turbofan Engine Remaining Useful Life. *FUDMA Journal of Sciences (FJS)*, 9(12), 581–589. <https://doi.org/10.33003/fjs-2025-0912-4237>
- Du, W., Zhang, J., Meng, G., & Zhang, H. (2024). Aero-engine fault detection with an LSTM auto-encoder combined with a self-attention mechanism. *Machines*, 12(12), 879. <https://doi.org/10.3390/machines12120879>
- Durand, J. G., Dubois, A., & Moss, R. J. (2023). Formal and Practical Elements for the Certification of Machine Learning Systems. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*. <https://doi.org/10.1109/DASC58513.2023.10311201>
- Ezhilarasu, C. M., Skaf, Z., & Jennions, I. K. (2021). A generalised methodology for the diagnosis of aircraft systems. *Ieee Access*, 9, 11437–11454. <https://doi.org/10.1109/ACCESS.2021.3050877>
- Huang, Y., Tao, J., Sun, G., Wu, T., Yu, L., & Zhao, X. (2023). A novel digital twin approach based on deep multimodal information fusion for aero-engine fault diagnosis. *Energy*, 270, 126894. <https://doi.org/10.1016/j.energy.2023.126894>
- Jia, J., & Li, Y. (2023). Deep learning for structural health monitoring: Data, algorithms, applications, challenges, and trends. *Sensors*, 23(21), 8824.
- Jia, L., Ezhilarasu, C. M., & Jennions, I. K. (2025). Fault Diagnosis Across Aircraft Systems Using Image Recognition and Transfer Learning. *Applied Sciences*, 15(6), 3232. <https://doi.org/10.3390/app15063232>
- Jia, Z., Li, Y., Wang, S., & Liu, Z. (2023). A multi-channel data-based fault diagnosis method integrating deep learning strategy for aircraft sensor system. *Measurement Science and Technology*, 34(2), 25115. <https://doi.org/10.1088/1361-6501/aca219>
- Kaakai, F., Dmitriev, K., Adibhatla, S., Baskaya, E., Bezecchi, E., Bharadwaj, R., Brown, B., Gentile, G., Gingins, C., & Grihon, S. (2022). Toward a machine learning development lifecycle for product certification and approval in aviation. *SAE International Journal of Aerospace* 339791, 15(2), 127–143. <https://doi.org/10.4271/01-15-02-0009>
- Kabashkin, I. (2024). Integration of Foundation Models and Federated Learning in AIoT-Based Aircraft Health Monitoring Systems. *Mathematics*, 12(21), 3428. <https://doi.org/10.3390/math12213428>
- Kabashkin, I. (2025). AI and Evolutionary Computation for Intelligent Aviation Health Monitoring. *Electronics* (2079-9292), 14(7). <https://doi.org/10.3390/electronics14071369>
- Kabashkin, I., Perekestov, V., & Pivovar, M. (2025). AI-Driven Fault Detection and Maintenance Optimization for Aviation Technical Support Systems. *Processes*, 13(3), 666. <https://doi.org/10.3390/pr13030666>
- Kefalas, M., Santiago Rojo Jr, J. de, Apostolidis, A., Van Den Herik, D., van Stein, B., & Bäck, T. (2022). Explainable artificial intelligence for exhaust gas temperature of turbofan engines. *Journal of Aerospace Information Systems*, 19(6), 447–454.
- Khanzada, H. R., Maqsood, A., & Basit, A. (2025). Artificial intelligence in UAV flight controls: deep reinforcement learning based altitude-hold strategies for fixed-wing UAVs. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2025.3581966>
- Kraemer, A. D., & Villani, E. (2024). Framework for Offline Data-Driven Aircraft Fault Diagnosis. *Journal of Aerospace Information Systems*, 21(4), 14. <https://doi.org/10.2514/1.I011253>
- Kwakye, A. D., Jennions, I. K., & Ezhilarasu, C. M. (2024). Platform health management for aircraft maintenance—a review. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 238(3), 267–283. <https://doi.org/10.1177/09544100231222516>

- Lei, B., Shen, N., Ji, D., Ouyang, R., Yang, L., Liu, Q., Gao, B., & Wang, Z. (2024). A comprehensive review of key technologies for enhancing the reliability of aero-engines. *IEEE Access*, 12, 148315–148331. <https://doi.org/10.1109/ACCESS.2024.3474737>
- Li, J., King, S., & Jennions, I. (2023). Intelligent fault diagnosis of an aircraft fuel system using machine learning—A literature review. *Machines*, 11(4), 481. <https://doi.org/10.3390/machines11040481>
- Li, J., King, S., & Jennions, I. (2025). Intelligent Multi-Fault Diagnosis for a Simplified Aircraft Fuel System. *Algorithms*, 18(2), 73. <https://doi.org/10.3390/a18020073>
- Li, M., Zhu, L., Luo, M., & Ke, T. (2025). Remaining Useful Life Prediction of Airplane Engine Based on Bidirectional Mamba and Causal Discovery. *Sensors*, 25(11), 3429. <https://doi.org/10.3390/s25113429>
- Li, Z., Ma, J., Fan, R., Zhao, Y., Ai, J., & Dong, Y. (2025). Aircraft Sensor Fault Diagnosis Based on GraphSage and Attention Mechanism. *Sensors (Basel, Switzerland)*, 25(3), 809. <https://doi.org/10.3390/s25030809>
- Li, Z., Zhang, Y., Ai, J., Zhao, Y., Yu, Y., & Dong, Y. (2023). A lightweight and explainable data-driven scheme for fault detection of aerospace sensors. *IEEE Transactions on Aerospace and Electronic Systems*, 59(6), 8392–8410. <https://doi.org/10.1109/TAES.2023.3303855>
- Liu, Y., Dong, J., Li, Y., Gong, X., & Wang, J. (2022). A UAV-based aircraft surface defect inspection system via external constraints and deep learning. *IEEE Transactions on Instrumentation and Measurement*, 71, 1–15. <https://doi.org/10.1109/TIM.2022.3198713>
- Luettig, B., Akhiat, Y., & Daw, Z. (2024). ML meets aerospace: challenges of certifying airborne AI. *Frontiers in Aerospace Engineering*, 3, 1475139. <https://doi.org/10.3389/fpace.2024.1475139>
- Ofoegbu, E., & Josiah, P. (2026). A review of artificial intelligence applications for fault detection in aerospace hydraulics systems. *Australian Journal of Multi-Disciplinary Engineering*, 1–17. <https://doi.org/10.1080/14488388.2026.2621623>
- Plevris, V., & Papazafeiropoulos, G. (2024). AI in structural health monitoring for infrastructure maintenance and safety. *Infrastructures*, 9(12), 225. <https://doi.org/10.3390/infrastructures9120225>
- Shafi, I., Mazhar, M. F., Fatima, A., Alvarez, R. M., Miró, Y., Espinosa, J. C. M., & Ashraf, I. (2023). Deep learning-based real time defect detection for optimization of aircraft manufacturing and control performance. *Drones*, 7(1), 31. <https://doi.org/10.3390/drones7010031>
- Shen, Y., & Khorasani, K. (2020). Hybrid multi-mode machine learning-based fault diagnosis strategies with application to aircraft gas turbine engines. *Neural Networks*, 130, 126–142. <https://doi.org/10.1016/j.neunet.2020.07.001>
- Sridhar, C., Gupta, V., Jain, P., & Vaidhyanathan, K. (2025). Approach towards semi-automated certification of low criticality ML-enabled airborne applications. *2025 IEEE/ACM 4th International Conference on AI Engineering—Software Engineering for AI (CAIN)*, 107–112.
- Tsallis, C., Papageorgas, P., Piromalis, D., & Munteanu, R. A. (2025). Application-wise review of Machine Learning-based predictive maintenance: Trends, challenges, and future directions. *Applied Sciences*, 15(9), 4898. <https://doi.org/10.3390/app15094898>
- Wu, J., Kong, L., Kang, S., Zuo, H., Yang, Y., & Cheng, Z. (2024). Aircraft engine fault diagnosis model based on 1DCNN-BiLSTM with CBAM. *Sensors*, 24(3), 780. <https://doi.org/10.3390/s24030780>
- Young, L. K., Hussain, K., & Oh, I.-Y. (2025). Validation for Predictive Maintenance of Aircraft Systems Using AI Models Developed With Rotor Blade Vibration Data. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2025.3550822>
- Zhang, W., Liu, Z., Jia, Z., Wang, X., Yan, W., & Wang, K. (2025). Application of a multimodal deep learning model based on recursive fusion feature map with transformer-TCN for complex fault diagnosis of flying wing UAV actuators. *IEEE Transactions on Instrumentation and Measurement*. <https://doi.org/10.1109/TIM.2025.3561438>

- Zhao, Y.-P., & Chen, Y.-B. (2022). Extreme learning machine based transfer learning for aero engine fault diagnosis. *Aerospace Science and Technology*, 121, 107311. <https://doi.org/10.1016/j.ast.2021.107311>
- Zhao, Y., Zhao, H., Ai, J., & Dong, Y. (2022). Robust Data-Driven Fault Detection: An Application to Aircraft Air Data Sensors. *International Journal of Aerospace Engineering*, 2022(1), 2918458.
- Zolghadri, A. (2024). A review of fault management issues in aircraft systems: Current status and future directions. *Progress in Aerospace Sciences*, 147, 101008. <https://doi.org/10.1016/j.paerosci.2024.101123>