Edge-Deployed Generative Language-Based Retrieval for Aerospace Asset Health Management

Lukmon Rasaq[1], Madhuri Siddula[2], Om Prakash Yadav[3], Rhonda Walthall[4], Joseph Ensberg[5], Piyush Yadav[6]

[1,2,3]*North Carolina A&T State University, Greensboro, NC, 27411, USA*
{larasaq@aggies.ncat.edu, msiddula@ncat.edu, oyadav@ncat.edu}

[4,5]*Collins Aerospace, Charlotte, NC, 28217, USA*

[6]*Collins Aerospace, Cork, Ireland*
{rhonda.walthall@collins.com, joseph.j.ensberg@rtx.com, Piyush.Yadav@collins.com}

ABSTRACT

Aerospace asset health management increasingly relies on access to large volumes of maintenance documentation; however, operational environments are often constrained by limited connectivity and computational resources, restricting the use of cloud-based intelligence systems. This paper presents a fully offline, edge-deployable retrieval-augmented generation (RAG) framework for aerospace maintenance and prognostics using technical documentation. The framework integrates a locally hosted lightweight large language model with vector retrieval and cross-encoder reranking to support natural language querying of Airworthiness Directives (ADs) and maintenance records. Deployed on an NVIDIA Jetson Orin Nano 8GB device, the system performs document ingestion, indexing, retrieval, reranking, and response generation entirely on-device without cloud connectivity. Experimental evaluation on real aerospace maintenance documents demonstrates the ability to identify failure mechanisms, failure modes, root causes, affected components, and maintenance procedures described in FAA documentation. The cross-encoder reranking stage improves retrieval precision by refining semantically overlapping maintenance evidence prior to generation. The framework achieved an average fidelity score of 0.83, indicating that most generated responses remained grounded in retrieved FAA evidence. Across representative AD queries, the system achieved practical edge inference latency of approximately 4–10 seconds on the Jetson platform. The results demonstrate the feasibility of privacy-preserving, low-latency generative artificial intelligence for aerospace maintenance decision support on resource-constrained edge devices.

1. INTRODUCTION

Modern aircraft systems operate in highly regulated, business-competitive environments where maintenance, accuracy, reliability, and timely decision-making are essential. As aircraft architecture grows increasingly complex and maintenance documentation continues to expand, aerospace personnel must interpret large volumes of technical and regulatory material to diagnose faults and ensure compliance. The integration of artificial intelligence into maintenance workflows presents an opportunity to enhance document understanding, improve decision support, and strengthen asset health management. However, deploying intelligent systems in operational aerospace environments requires addressing challenges related to reliability, privacy, computational constraints, and offline execution.

1.1. Role of AI and PHM in Aircraft Maintenance

The reliability and availability of modern aircraft depend on effective maintenance strategies enabled by prognostics and health management (PHM) frameworks. PHM supports asset condition monitoring, fault diagnosis, degradation prediction, and remaining useful life estimation, thereby enabling condition-based maintenance and reducing unplanned downtime (Fu & Avdelidis, 2023). In parallel, engineering asset management integrates reliability, availability, and maintainability principles with PHM to support lifecycle-oriented decision-making for complex industrial systems (Payette & Abdul Nour, 2023).

Aircraft maintenance operations are highly knowledge-intensive and rely on extensive technical documentation, including Aircraft Maintenance Manuals, Fault Isolation Manuals, Airworthiness Directives, and regulatory guidance. Maintenance personnel must efficiently interpret these documents to diagnose faults, perform troubleshooting, and ensure compliance with prescribed maintenance procedures.

However, manual search across large and heterogeneous aerospace documentation repositories can be time-consuming and operationally inefficient (Lei Hou, Beixi Jia, Chenguang Xing, Zhaojiang Chen, Ziliang Du, 2025). Furthermore, traditional keyword-based retrieval methods often struggle to provide accurate information discovery across distributed aerospace knowledge bases (Fei Chen, Zhonghua Wen, Bo Liu, 2025). These limitations motivate the adoption of artificial intelligence techniques to improve document understanding and maintenance decision support in aerospace environments.

1.2. Large Language Models and Retrieval-Augmented Generation

Recent advances in large language models (LLMs) have enabled powerful natural language understanding and generation capabilities, creating new opportunities for intelligent question answering and maintenance assistance. However, general-purpose LLMs are trained on static corpora and may hallucinate when answering domain-specific questions, particularly in specialized fields such as aerospace engineering (Chen et al., 2025). To address this limitation, Retrieval Augmented Generation (RAG) has emerged as an effective approach that combines LLM reasoning with external knowledge sources by retrieving relevant documents during inference (Yadav, 2024). By grounding model responses in retrieved documentation, RAG frameworks can improve answer relevance, contextual accuracy, and traceability. Recent aerospace maintenance assistant systems have demonstrated that integrating domain-specific document repositories with vector-based retrieval improves the effectiveness of question answering workflows for maintenance documentation (Hou et al., 2025). In such systems, technical documents are encoded into dense vector representations that enable semantic search across large document collections. This architecture allows maintenance engineers to query complex technical documentation using natural language while retrieving relevant procedural guidance directly from authoritative sources.

1.3. Challenges of RAG at the Edge During Aircraft Maintenance and Offline Phases

Despite these advances, deploying RAG systems in operational environments remains challenging, particularly on resource-constrained edge devices. RAG introduces substantial memory and computation overhead due to vector similarity search and embedding storage, which can exceed the memory capacity of mobile and embedded platforms (Korakit Seemakhupt, Sihang Liu, Samira Khan 2024). Recent research has proposed techniques such as online indexed retrieval and adaptive embedding caching to reduce memory footprint and latency for edge-based RAG systems, demonstrating feasibility on platforms such as the NVIDIA Jetson Orin Nano (Seemakhupt et al., 2024). In addition, domain-specific RAG methods that emphasize efficient ranking of retrieved documents rather than complex reasoning have shown improved suitability for small-scale LLMs operating in edge environments (Juntae Lee, Jihwan Bang, Kyuhong Shim, Seunghan Yang, Simyung Chang, 2025). However, most existing aerospace RAG and LLM-based systems focus on cloud or workstation deployment and do not explicitly address fully offline operation, maintenance document handling, and secure execution in field environments. Moreover, while prior work demonstrates intelligent question answering, limited attention has been given to integrating such systems directly into PHM-oriented workflows that emphasize the extraction of failure mechanisms, failure modes, root causes, affected locations, and prescribed maintenance procedures from regulatory and technical documents.

1.4. Contribution

Motivated by these gaps, this paper presents a fully offline, edge-deployed generative language-based retrieval framework designed to support aerospace asset health management using maintenance documentation. The primary contributions of this work are as follows:

1. Development of a two-stage dense retrieval and cross-encoder reranking framework that improves semantic grounding, reduces cross-document blending, and enhances retrieval precision for aerospace maintenance question answering on resource-constrained edge devices.
2. Design and implementation of a fully offline RAG framework deployed on an NVIDIA Jetson Orin Nano 8GB device, enabling document ingestion, indexing, retrieval, reranking, and response generation entirely on the device without cloud dependency.
3. Demonstration of PHM-oriented intelligence extraction, including identification of failure mechanisms, failure modes, root causes, affected system locations, and prescribed maintenance procedures from Airworthiness Directives and OEM documentation.

By executing all processing stages locally, the proposed framework supports privacy preservation, secure deployment in constrained environments, and practical decision support for aerospace asset health management. The remainder of the paper is organized as follows. Section 2 presents the system methodology, including system overview (Section 2.1), document collection and PDF parsing (Section 2.2), token aware text segmentation and embedding generation (Section 2.3), FAISS vector indexing and similarity-based retrieval (Section 2.4), two-stage dense retrieval and cross-encoder reranking (Section 2.5), large language model inference and response generation (Section 2.6), and evaluation metrics (Section 2.7). Section 3 describes the experimental setup,

Section 4 presents the results, Section 5 discusses the findings, and Section 6 concludes the paper.

2. RESEARCH APPROACH AND METHODS

In operational aerospace environments, maintenance activities frequently occur in memory and compute-constrained or security-sensitive settings where cloud connectivity is limited or unavailable. Consider a maintenance technician diagnosing a fault condition referenced in an Airworthiness Directive (AD) while the aircraft is positioned in a remote hangar. The technician must rapidly identify the failure mechanism, affected component location, and mandated corrective action from extensive regulatory documentation. Traditional document search workflows require manual navigation across multiple PDF manuals and directives, which can be time-consuming and prone to oversight. In aerospace contexts, delayed or inaccurate interpretation may impact turnaround time, regulatory compliance, and asset reliability. These operational constraints motivate the development of a secure, fully offline intelligent retrieval system capable of delivering authoritative, evidence-grounded maintenance guidance directly on edge hardware.

This paper adopts a system-oriented, application-driven approach to design, implement, and evaluate a fully offline, language-based retrieval framework for aerospace asset health management under edge-computing constraints. The primary objective is to demonstrate that a large language model powered by a retrieval-augmented generation framework can be deployed locally on resource-constrained platforms to provide reliable, privacy-preserving, and practical maintenance intelligence from aerospace documentation (e.g., AD).

2.1. System Overview

Figure 1 illustrates a fully offline, edge-deployed Retrieval-Augmented Generation (RAG) framework implemented on an NVIDIA Jetson Orin Nano 8 GB for aerospace maintenance assistance. Aerospace maintenance PDFs are parsed into document segments, embedded using a lightweight bi-encoder model, and indexed locally within a FAISS vector database. At query time, a user submits a maintenance question via a Streamlit interface, and semantically similar candidates (Top-k_1) are retrieved through vector similarity search. A cross-encoder relevance reranking model further refines these candidates to produce a smaller set of high-confidence evidence corresponding to authoritative corrective actions (Top-k_2). The selected evidence is then supplied as contextual input to a locally served Llama 3.2 3B model via Ollama, which generates an evidence-grounded response. The lower portion of the figure depicts the high-level device initialization steps required for edge deployment on the Jetson Orin Nano 8 GB, summarizing the JetPack-based deployment initialization process used to prepare the runtime environment for on-device inference.

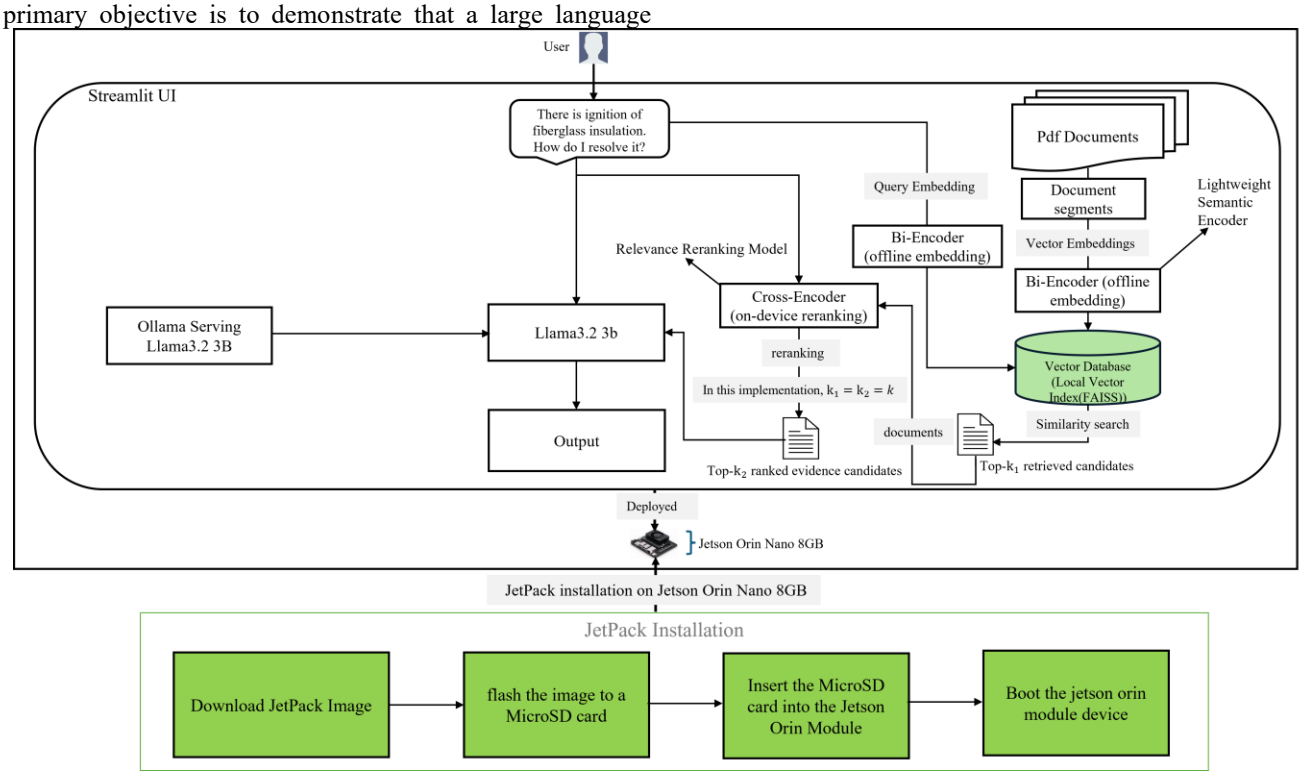

Figure 1. Fully offline, edge-deployed RAG framework on Jetson Orin Nano 8 GB, combining FAISS-based retrieval, cross-encoder reranking, and on-device LLM generation for aerospace maintenance assistance.

2.2. Document Collection and PDF Parsing

A curated corpus of aerospace maintenance documentation is assembled to evaluate the proposed retrieval-augmented generation framework. Initially, OEM manuals and procedural maintenance records are used to verify baseline system functionality and confirm correct document ingestion, indexing, and retrieval behavior. These documents provide structured descriptions of aircraft subsystems, troubleshooting steps, and prescribed maintenance actions that reflect real operational workflows. Subsequently, fifty Airworthiness Directives (ADs) are collected from the Federal Aviation Administration database and incorporated into the knowledge base. These ADs contain authoritative regulatory information describing unsafe conditions, affected components, failure modes, root causes, locations within the aircraft where failures occur, and mandated maintenance procedures required to address identified issues (Federal Aviation Administration, 2024; Lukmon Rasaq, Korbin Ferguson, William Teasley, Max Xu, Kyle E. Blond, Om Prakash Yadav, 2024). The inclusion of ADs enables evaluation of the framework on operational regulatory content, which is central to aerospace asset health management. Documents are uploaded through a Streamlit-based user interface and parsed locally using a PDF loader utility, which converts each file into page-level text representations while preserving source metadata. The parsed pages are then forwarded to the segmentation stage for downstream chunking and embedding. All document ingestion and parsing operations are performed entirely on the local device, ensuring offline operation and preventing proprietary maintenance data from leaving the execution environment.

2.3. Token-Aware Text Segmentation and Embedding Generation

After PDF parsing and preprocessing, each document is divided into smaller text segments to balance semantic granularity and contextual completeness. Prior aerospace question answering and maintenance assistant systems have shown that unstructured technical documents must be partitioned into multiple text chunks before downstream representation learning and retrieval to support effective semantic matching and scalable storage (Chen et al., 2025). Similarly, recent RAG frameworks perform indexing by splitting incoming data into smaller overlapping chunks before embedding generation (Seemakhupt et al., 2024). In the proposed framework, segmentation is implemented using a token-aware recursive text splitter driven by a Hugging Face tokenizer associated with the embedding model. The overall token-aware segmentation and embedding pipeline is illustrated in Figure 2.

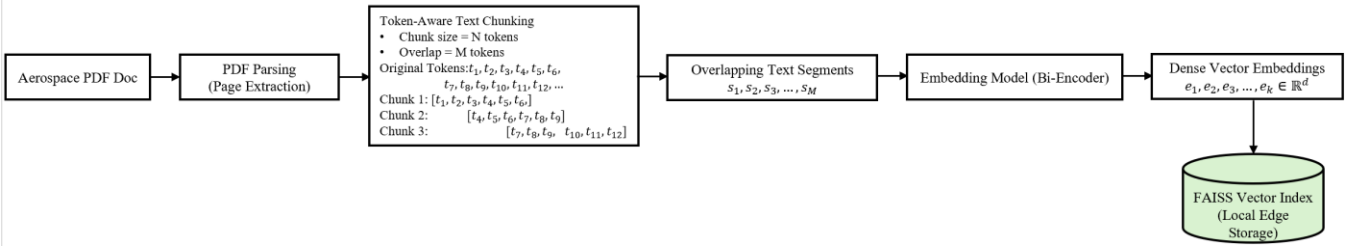

Figure 2. Token-aware segmentation and embedding pipeline for offline index construction. Aerospace PDF documents are parsed and partitioned into overlapping token-based segments, encoded into dense vectors $\boldsymbol{e}_i \in \mathbb{R}^d$ using a bi-encoder model, and stored locally in a FAISS index to enable efficient similarity-based retrieval.

As shown in Figure 2, aerospace PDF documents are parsed and segmented into overlapping token-based chunks, which are encoded into dense vector embeddings and indexed locally using FAISS. This preprocessing stage enables efficient similarity-based retrieval while preserving contextual continuity through controlled token overlap.

Let a document be represented as a sequence of tokens:

$$D = [w_1, w_2, w_3, \ldots, w_N] \tag{1}$$

where w_i denotes the i-th token and N is the document length. The document is partitioned into a set of overlapping segments:

$$S = \{s_1, s_2, s_3, \ldots, s_M\} \tag{2}$$

where each segment $s_j \in D$ consists of a fixed-length token sequence with controlled overlap to preserve contextual continuity. Overlapping chunking is commonly used in RAG pipelines to preserve continuity across segment boundaries (Seemakhupt et al., 2024). Each segment is mapped to a dense vector representation using a pretrained sentence-embedding model. Let $f(\cdot)$ denote the embedding function. The embedding of the segment s_j is computed as:

$$\boldsymbol{e}_j = f(s_j), \quad \boldsymbol{e}_j \in \mathbb{R}^d \tag{3}$$

where d is the embedding dimensionality and $\boldsymbol{e}_j$ represents the dense vector encoding of the document segment (Reimers & Gurevych, 2019; Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih, 2020). Similar

vectorization strategies are widely used in aerospace RAG systems to capture semantic relationships among technical concepts (Hou et al., 2025). During inference, the user query $q$is encoded using the same embedding function:

$$\boldsymbol{e}_q = f(q),\ \boldsymbol{e}_q \in \mathbb{R}^d \tag{4}$$

Following standard bi-encoder dense retrieval, document segments and queries are represented in a shared embedding space to enable similarity-based retrieval (Karpukhin et al., 2020). Moreover, semantic similarity between the query embedding and each segment embedding is computed using dot product similarity, which is the metric implemented in the FAISS vector store within the proposed system (J. Johnson, M. Douze, H. Jégou, 2019). The dot product similarity is defined as:

$$\mathrm{sim}\left(\boldsymbol{e}_q, \boldsymbol{e}_j\right) = \boldsymbol{e}_q^{\mathrm{T}}\, \boldsymbol{e}_j \tag{5}$$

Additionally, segments with the highest similarity scores are selected as candidate evidence for retrieval. This embedding-based representation enables semantic matching beyond keyword overlap and forms the foundation for accurate retrieval in the proposed framework, consistent with established RAG indexing and lookup pipelines (Seemakhupt et al., 2024).

2.4. FAISS Vector Indexing and Similarity-Based Retrieval

After embedding generation, all document segment embeddings are indexed to enable efficient similarity-based retrieval. Let $E = \{\mathbf{e}_j\}_{j=1}^{k_2}$ denote the set of embedding vectors corresponding to the segmented document corpus, where each $\mathbf{e}_j$ is computed from its associated text segment s_j according to Eq. (3). Dense retrieval methods represent documents and queries within a shared vector space and perform similarity search to identify semantically related content (Karpukhin et al., 2020).

To support efficient retrieval on resource-constrained edge devices, the proposed framework employs Facebook AI Similarity Search (FAISS), a high-performance vector indexing and similarity search library developed by Facebook AI Research (Johnson et al., 2019). During inference, the query embedding $\mathbf{e}_q$ is compared against the indexed vectors to retrieve the top-k most semantically relevant document segments:

$$N_k(\mathbf{e}_q) = \underset{k,\, \mathbf{e}_j \in E}{\arg\operatorname{top}}\ \mathrm{sim}(\mathbf{e}_q, \mathbf{e}_j) \tag{6}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the dot-product similarity function defined in Eq. (5). This formulation is consistent with dense passage retrieval frameworks (Karpukhin et al., 2020). FAISS enables efficient approximate nearest neighbor retrieval while maintaining high recall, making it suitable for edge-oriented RAG deployment on platforms such as the NVIDIA Jetson Orin Nano (Johnson et al., 2019; Seemakhupt et al., 2024).

The retrieved top-k segments are subsequently forwarded to the reranking and context construction stage, where their relevance is further refined before being provided as contextual input to the local language model. This two-stage retrieval strategy balances retrieval efficiency and semantic accuracy while remaining suitable for fully offline edge deployment.

2.5. Two-Stage Dense Retrieval and Cross-Encoder Reranking

After similarity-based retrieval using a bi-encoder embedding model, the retrieved top-k document segments may still contain partially redundant or weakly relevant content. Although bi-encoders enable efficient retrieval by independently embedding queries and documents into a shared vector space, they do not explicitly model fine-grained interactions between query and document text, which can limit ranking precision (Karpukhin et al., 2020). This limitation is particularly important in aerospace maintenance corpora, where semantically related Airworthiness Directives often contain overlapping subsystem terminology and corrective procedures.

To improve semantic grounding and reduce semantic retrieval noise, a cross-encoder reranking stage is applied to the candidate segments returned by the bi-encoder. Unlike bi-encoders, a cross-encoder jointly processes the query and each candidate segment as a single input sequence, enabling deeper contextual interaction modeling and more accurate relevance estimation (Nogueira and Cho, 2019; Karpukhin et al., 2020). Let $N_k\left(\boldsymbol{e}_q\right) = \{s_{j_1}, s_{j_2}, s_{j_3} \ldots, s_{j_k}\}$ denote the set of top-k candidate segments returned by the retrieval stage. The cross-encoder computes a relevance score for each query-segment pair:

$$r_i = g\left(q, s_{j_i}\right) \tag{7}$$

where $g(\cdot,\cdot)$ denotes the cross-encoder scoring function. The candidate segments are then reordered according to r_i. In this work, the retrieval size and reranking size are set equal ($k_1 = k_2 = k$), such that all retrieved candidates are reranked and retained for context construction. This two-stage retrieval strategy improves retrieval precision and semantic grounding while maintaining practical latency for offline edge deployment (Nogueira & Cho, 2019; Seemakhupt et al., 2024). The reranked segments are concatenated in ranked order to construct the contextual prompt supplied to the local language model:

$$C = \text{Concat}\left(s_{j_{(1)}}, s_{j_{(2)}}, \cdots s_{j_{(k_2)}}\right) \quad (8)$$

where Concat $(\cdot)$ represents ordered sequence concatenation. This context construction strategy follows the retrieval-augmented generation paradigm, where reranked passages are combined to form the conditioning context for language model inference (Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela, 2020).

2.6. Large Language Model Inference and Response Generation

Large Language Models (LLMs) exhibit strong natural language understanding capabilities but may hallucinate when queried about domain-specific or unseen information. Retrieval-Augmented Generation (RAG) mitigates this limitation by conditioning language model outputs on externally retrieved evidence rather than relying solely on internal model parameters (Lewis et al., 2020). In the proposed framework, the final reranked evidence set constructed in Eq. (8) is supplied as contextual input to a locally hosted small-scale LLM. Let C denote the concatenated evidence context formed from the top- k reranked segments and let q denote the user query. The generation process is defined as:

$$y = \mathcal{L}(C, q) \quad (9)$$

where $\mathcal{L}(\cdot)$ represents the local language model and y is the generated response. The context is supplied prior to the query to prioritize evidence grounding during generation. The framework employs Llama 3.2 3B served locally via Ollama. Instead of relying on model quantization, computational efficiency is achieved using a compact model that balances reasoning capability and resource consumption, enabling deployment on the NVIDIA Jetson Orin Nano 8GB while maintaining acceptable response quality. The model is instructed to generate responses strictly from the supplied context C, reducing hallucination and improving factual consistency for aerospace maintenance applications. Similar RAG-based maintenance assistants have shown that grounding generation in vectorized technical knowledge bases improves answer relevance and reliability (Hou et al., 2025).

Unlike prior aerospace RAG systems such as (Hou et al., 2025) and (Yadav, 2024), which primarily target centralized computing environments, the proposed framework is designed for fully offline edge deployment. All stages of ingestion, indexing, retrieval, reranking, and generation are executed locally without cloud connectivity, enabling deployment in bandwidth-constrained and security-sensitive maintenance environments. Recent edge-oriented RAG studies further indicate that retrieval operations constitute the dominant system overhead rather than language generation itself (Seemakhupt et al., 2024). Accordingly, the framework employs FAISS-based approximate nearest neighbor search and a two-stage retrieval strategy to maintain practical latency while preserving retrieval precision. Lightweight ranking methods have similarly been shown to improve efficiency for domain-specific RAG on edge devices (Lee et al., 2025). This final stage transforms reranked technical evidence into grounded natural language responses for aerospace maintenance decision support.

2.7. Evaluation Metrics

The evaluation metrics used in this work are fidelity and latency. Fidelity evaluates the faithfulness of the generated response to the retrieved aerospace maintenance documentation. Following prior RAG-based maintenance evaluation approaches, fidelity is defined as the ratio between the number of generated statements supported by retrieved FAA Airworthiness Directive evidence and the total number of generated statements (Hou et al., 2025).

$$F = \frac{|V|}{|S|} \quad (10)$$

where V represents the number of generated statements verified from the retrieved FAA Airworthiness Directive evidence and S represents the total number of generated statements in the response. A fidelity score of 1.0 indicates fully grounded maintenance guidance, while lower scores indicate partial grounding or semantic blending across related Airworthiness Directives.

Latency evaluates the end-to-end response time from query submission to answer generation, including query embedding computation, vector retrieval, cross-encoder reranking, and local language model inference. This metric reflects the practical responsiveness of the proposed framework when deployed on resource-constrained edge hardware.

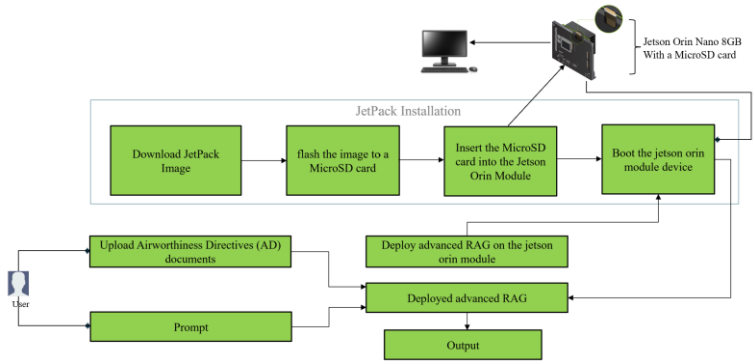

Figure 3. JetPack-based edge deployment workflow for the fully offline aerospace RAG framework on the NVIDIA Jetson Orin Nano 8 GB.

Additionally, Figure 3 depicts the end-to-end deployment workflow of the proposed aerospace RAG system on an NVIDIA Jetson Orin Nano 8 GB. It begins with JetPack image installation, where the operating system image is flashed onto a MicroSD card, inserted into the Jetson Orin module, and the device is booted to initialize the edge runtime environment. The RAG framework is then deployed locally, and FAA Airworthiness Directives (ADs) are uploaded, parsed, embedded, and indexed entirely on the device. During operation, user prompts are fully processed on the Jetson platform, with retrieval, reranking, and response generation executed locally to provide maintenance decision support without relying on cloud connectivity.

Furthermore, Algorithm 1 presents a fully offline, edge-deployed RAG pipeline designed for aerospace maintenance assistance. During offline ingestion, aerospace documents are parsed and divided into token-aware, overlapping segments using SPLIT-INTO-SEGMENTS(·) to preserve semantic continuity. These segments are embedded using the semantic encoder $f(\cdot)$ and indexed via BUILD-FAISS-INDEX(·) to enable efficient dense similarity search. At query time, the user query is embedded to obtain $\boldsymbol{e}_q$, and TOP- k_1 -SIMILAR(·) retrieves the Top- k_1 semantically similar segments from the FAISS index. A cross-encoder relevance reranking model assigns relevance scores to these candidates, and TOP- k_2 -BY-SCORE(·) selects the most informative evidence segments. The selected segments are cleaned and concatenated using CONCAT(CLEAN-TEXT(·)) to form a compact, context-restricted prompt. The constructed context is then provided to a locally deployed language model, which generates an evidence-grounded response suitable for low-latency execution on resource-constrained edge platforms.

Additionally, in Algorithm 1, e_q denotes the embedding of the user query q. In the implementation, the raw query q is supplied directly to the retrieval function, which internally computes e_q using the embedding model $f(\cdot)$.

ALGORITHM 1-FULLY AEROSPACE EDGE DEPLOYED PIPELINE

function AEROSPACE-RAG(D, q, f, I, g, k)
D: document set; q: user query; f: embedding model
I: FAISS index; g: cross-encoder reranker; k: retrieval size
▷ Offline ingestion and indexing
CORPUS ← ∅
for each document $D^{(i)} \in D$ **do**
▷ PDF parsing and segmentation
PAGES ← LOAD-PDF-PAGES($D^{(i)}$)
▷ token-aware splitter, overlap
$S^{(i)}$ ← SPLIT-INTO-SEGMENTS(PAGES)
CORPUS ← CORPUS ∪ $S^{(i)}$
end for
I ← BUILD-FAISS-INDEX(CORPUS, f)
▷ Encode user query
$e_q \leftarrow f(q)$
▷ Retrieve top-k candidate segments from FAISS (dot-product similarity)
$N_k(e_q) \leftarrow$ TOP_k_SIMILAR(I, e_q, k)
▷ Cross-encoder reranking over retrieved candidates
for each segment $s_{j_i} \in N_k(e_q)$ **do**
$r_i \leftarrow g(q, s_{j_i})$ ▷ Compute rerank score
end for
▷ Rerank retrieved candidates and retain top-k ($k_1 = k_2 = k$)
$\mathcal{R}_k \leftarrow$ TOP_k_BY_SCORE $\left(\left(\{s_{j_i}, r_i\}\right),\ k\right)$
▷ Build cleaned context from reranked segments
$C \leftarrow$ CONCAT(CLEAN_TEXT(s) for $s \in \mathcal{R}_k$)
▷ Local LLM generation (Ollama)
y ← LLM-GENERATE(C, q)
return y
end function

3. EXPERIMENT

Experiments were conducted on an NVIDIA Jetson Orin Nano 8 GB edge device running JetPack on Ubuntu Linux to evaluate the proposed fully offline Retrieval-Augmented Generation (RAG) framework for aerospace maintenance assistance. The system was implemented using a Streamlit-based interface, a lightweight sentence-transformer embedding model (all-MiniLM-L12-v2), a FAISS vector database with dot-product similarity, and a cross-encoder reranker (ms-marco-TinyBERT-L-2-v2). Multiple language models, including Llama 2 7B chat-hf (Touvron et al., 2023), Llama 3.2 3B, Facebook BlenderBot, and Mistral 7B Instruct-v2.0, were evaluated for edge deployment. Due to the limited shared memory of the Jetson Orin Nano 8 GB device, larger models such as Mistral 7B Instruct-v2.0 resulted in system instability. Consequently, Llama 3.2 3B was selected, as it provided stable execution with approximately 3 GB memory usage during inference while maintaining acceptable response quality. Response generation was performed using this model deployed locally via Ollama.

A local corpus of aerospace maintenance documentation, including OEM manuals and Federal Aviation Administration (FAA) Airworthiness Directives (ADs), was processed entirely on-device. OEM and maintenance documents were initially used to validate system functionality, after which fifty FAA Airworthiness Directives were incorporated for evaluation. All documents were parsed, segmented into overlapping text segments, embedded, and indexed locally. During inference, user queries were encoded, Top- k_1 candidate segments were retrieved via dense similarity search, and subsequently reranked using a cross-encoder. The Top- k_2 reranked segments were concatenated to form a contextual prompt, which was supplied to the locally deployed language model. All retrieval and generation operations were executed fully offline on the edge device.

The framework was evaluated using real-world aircraft maintenance documentation to verify that generated responses were consistent with the information contained in

the Airworthiness Directives. System outputs were compared against reference maintenance information extracted directly from the documents. By grounding responses in FAA Airworthiness Directives, the system enables accurate identification of failure mechanisms, failure modes, affected components, and mandated corrective actions described in the maintenance documentation. The combined dense retrieval and cross-encoder reranking pipeline helps ensure that the language model receives relevant contextual evidence during response generation.

4. RESULTS

Table 1 summarizes representative query–response examples together with the corresponding inference latency measured on the NVIDIA Jetson Orin Nano 8 GB edge device and a high-end laptop platform equipped with a 16 GB GPU. The results show that the proposed offline RAG framework generates responses that are generally consistent with the failure descriptions and corrective actions specified in the FAA Airworthiness Directives. In particular, the two-stage dense retrieval and cross-encoder reranking pipeline improves semantic grounding by refining the relevance ordering of retrieved maintenance segments before language model inference. This reranking stage is especially important in aerospace maintenance corpora, where semantically related Airworthiness Directives often contain overlapping subsystem terminology and corrective procedures. The framework successfully retrieves and extracts maintenance information such as failure mechanisms, failure modes, affected components, and corrective procedures, indicating that the reranked segments provide contextual grounding for the language model output.

Additionally, Figures 4 and 5 present the end-to-end inference latency per prompt for the Jetson Orin Nano and laptop platforms, respectively. As expected, the Jetson device exhibits higher latency due to limited computational resources and shared memory constraints. Nevertheless, most responses on the edge device are generated within approximately 4–10 seconds, which remains suitable for interactive maintenance support scenarios. The framework achieved an average fidelity score of 0.83, demonstrating that most generated responses remained grounded in FAA maintenance documentation. These results further indicate that cross-encoder reranking improves retrieval precision and semantic grounding by reducing semantic retrieval noise prior to response generation. A notable observation is the elevated latency for Inference ID A on the Jetson device (40.64 seconds), which is attributed to a longer retrieved context and additional reranking overhead. In this case, the retrieved evidence segments contained more extensive textual content from the Airworthiness Directive, increasing both reranking computation and the context length supplied to the language model. Although one representative query exhibited reduced fidelity due to cross-document blending between semantically related maintenance directives, the reranking pipeline generally improved grounding consistency across the evaluated queries. Conversely, the laptop platform shows a lower latency for Inference ID E (3.55 seconds). This reduction is associated with a shorter contextual prompt and fewer retrieved segments required to answer the query. Since the retrieval and reranking stages produced a smaller final context in this case, the language model processed less input text and generated a response more quickly. Overall, the framework achieved grounded aerospace maintenance retrieval with practical edge inference latency. The findings further demonstrate that cross-encoder reranking plays a critical role in improving semantic grounding and reducing semantic retrieval noise within semantically overlapping aerospace maintenance corpora.

Table 1. The Commercial Aircraft Chatbot Performance on Fifty Uploaded Airworthiness Directives (ADs) Documents

AD Number	Inference ID	Prompt	RAG Response on Jetson Orin Nano 8GB	RAG Response on Laptop (16GB GPU)	Latency
2000-26-05	A	There is ignition of fiberglass insulation. How do I resolve it?	To resolve ignition of fiberglass insulation, install protective sleeving on the wiring, reroute the wire bundles, and repair any damaged wiring according to Boeing Alert Service Bulletin 767-35A0028, dated September 7, 1995.	To prevent potential ignition of fiberglass insulation in the environmental system (ECS) ducts, accomplish the following: Rework or Replacement-within 6 years after the effective date of this AD, rework ECS duct assemblies or replace existing duct assemblies with new or reworked duct assemblies, in accordance with Boeing Alert Service Bulletins.	Jetson Orin Nano 8GB: 40.64sec. Laptop: 7.63sec.
2000-26-05	B	Give me the estimated total cost for all the airplanes affected by the ignition of fiberglass insulation.	$2,552,996	The estimated total cost for all airplanes affected by this AD is $2,552,996.	Jetson Orin Nano 8GB: 4.29sec. Laptop: 8.54sec.
2004-23-14	C	What is the root cause of loss of DC power generation?	The root cause of loss of DC power generation was inadequate electrical ground path between the ground bracket of the transformer rectifier unit (TRU) and the structure, caused by improper preparation of the bonding surface.	Inadequate electrical ground path between the ground bracket of the left and right TRUs/MBC and the structure.	Jetson Orin Nano 8GB: 9.55sec. Laptop: 9.47sec.

2007-19-13	D	What causes unrestrained aft seatback rotation?	Under certain conditions, the Hydrolok retaining pin can migrate and disconnect from the seat structure.	Under certain conditions, the Hydrolok retaining pin can migrate and disconnect from the seat structure, resulting in the seat back having no rearward restraint and allowing it to rotate aft into the seat or exit pathway behind.	Jetson Orin Nano 8GB: 5.81sec. Laptop: 8.0sec.
2007-19-13	E	What action should I take when unrestrained aft seatback rotation occurs?	Reorient the Hydrolok retaining pin according to B/E Aerospace Alert Service Bulleting 25-20-2658	Visually inspect and re-orient if necessary, Hydrolok retaining pins within 30 days after the effective date of this AD	Jetson Orin Nano 8GB: 5.57sec. Laptop: 3.55sec.
6/18/2022	F	There is "RECIRC SMOKE" engine indication and crew alerting system (EICAS) messages.	The "RECIRC SMOKE" engine indication and crew alerting system (EICAS) messages may occur when water accumulates in the alternative ventilation system (AVS) duct.	The issue you're experiencing is due to water accumulating in the alternative ventilation system (AVS) duct. To resolve this, replace the alternative ventilation duct with a new part number as specified in the AD. Additionally, for certain airplanes, change the insulation blanket to install the drain hose.	Jetson Orin Nano 8GB: 7.06sec. Laptop: 4.01sec.

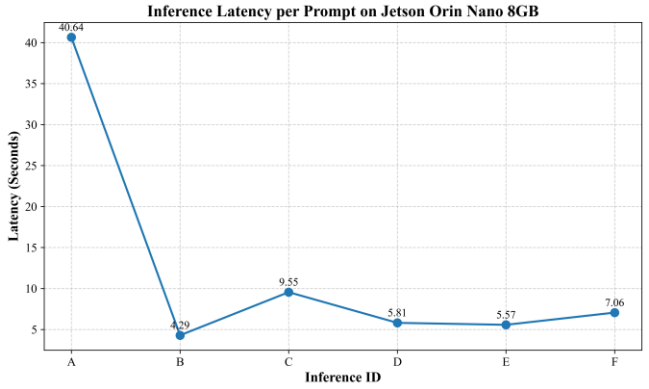

Figure 4. Inference latency per prompt for the proposed RAG framework using Llama 3.2 3B on the NVIDIA Jetson Orin Nano 8 GB.

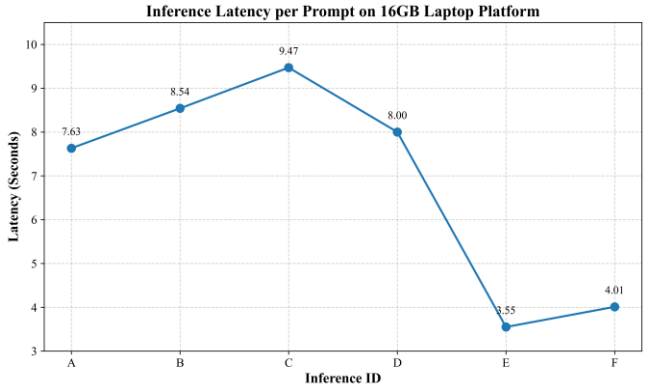

Figure 5. Inference latency per prompt for the proposed RAG framework using Llama 3.2 3B on a laptop with a 16 GB GPU.

5. DISCUSSION

The results demonstrate that a fully offline, edge-deployed RAG framework can provide evidence-grounded aerospace maintenance assistance using FAA Airworthiness Directives. The proposed framework reduces manual document search time by enabling direct natural-language querying of maintenance documentation, with most responses generated in approximately 4–10 seconds on the Jetson Orin Nano edge device. Although inference latency is higher on resource-constrained hardware, response times remain suitable for interactive maintenance scenarios. A key contribution of the framework is the two-stage dense retrieval and cross-encoder reranking pipeline, which improves semantic grounding by refining the relevance ordering of retrieved maintenance segments before language model inference. This reranking stage is particularly important in aerospace maintenance corpora, where semantically related Airworthiness Directives often contain overlapping subsystem terminology and corrective procedures. The improved contextual matching achieved through cross-encoder reranking reduces semantic retrieval noise and helps minimize irrelevant or misleading outputs during generation.

In one representative query involving fiberglass insulation ignition, the generated response combined corrective actions from multiple semantically related Airworthiness Directives. Although the retrieved evidence remained relevant to the query, the language model merged corrective procedures across related directives, thereby reducing response fidelity. This observation highlights the challenges of maintaining strict grounding consistency within semantically overlapping aerospace maintenance corpora, even when reranking improves retrieval precision. A key limitation of the current framework is its reliance on a compact language model and static document indexing, which may limit scalability as document volume and query complexity increase. Future work will focus on improving retrieval efficiency, evaluating additional edge-optimized language models, benchmarking alternative retrieval architectures and baseline retrieval pipelines without reranking, evaluating sensitivity to paraphrased maintenance queries, and integrating real-time sensor data to support tighter coupling with prognostics and health management workflows.

6. CONCLUSION

This paper presented a fully offline, edge-deployed Retrieval-Augmented Generation (RAG) framework for aerospace maintenance assistance designed to operate under strict computational, privacy, and connectivity constraints. By integrating dense retrieval, cross-encoder reranking, and a compact large language model, the framework generates evidence-grounded responses derived from authoritative aerospace documentation such as FAA Airworthiness Directives. The framework achieved an average fidelity score of 0.83 while maintaining practical edge inference latency. Experimental evaluation on an NVIDIA Jetson Orin Nano 8 GB demonstrated that the proposed approach achieves strong evidence grounding and acceptable latency for interactive use despite higher inference times relative to high-end hardware. These results confirm the feasibility of deploying reliable, privacy-preserving, and cloud-independent decision-support systems for aerospace asset health management on resource-constrained edge platforms. Collectively, the proposed framework establishes a practical foundation for next-generation intelligent maintenance systems operating in resource-constrained and security-sensitive environments.

REFERENCES

Fu, S., & Avdelidis, N. P. (2023). Prognostic and health management of critical aircraft systems and components: An overview. *Sensors*, *23*(19), 8124.

Payette, M., & Abdul-Nour, G. (2023). Asset management, reliability and prognostics modeling techniques. *Sustainability*, *15*(9), 7493.

Hou, L., Jia, B., Xing, C., Chen, Z., & Du, Z. (2025, February). Applied research on an aircraft maintenance assistant based on a large language model. In *Proceedings of the 2025 4th International Conference on Intelligent Systems, Communications and Computer Networks* (pp. 1-7).

Chen, F., Wen, Z., & Liu, B. (2025, June). A Question Answering System for Aerospace Large Language Models Based on Knowledge Graph and RAG Collaboration. In *Proceedings of the 2025 6th International Conference on Education, Knowledge and Information Management* (pp. 446-455).

Yadav, S. (2024, July). AeroQuery RAG and LLM for aerospace query in designs, development, standards, certifications. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1-6). IEEE.

Seemakhupt, K., Liu, S., & Khan, S. (2024). Edgerag: Online-indexed RAG for edge devices. *arXiv preprint arXiv:2412.21023*.

Lee, J., Bang, J., Shim, K., Yang, S., & Chang, S. (2025, April). Chain-of-rank: Enhancing large language models for domain-specific RAG in edge device. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 5601-5608).

Federal Aviation Administration. (2024). *Airworthiness Directives Manual*. Washington, DC: FAA.

Rasaq, L., Ferguson, K., Teasley, W., Xu, M., Blond, K. E., & Yadav, O. P. (2024, January). Sensor and Maintenance Strategy Evaluation for Boeing 767 Commercial Fleets. In *2024 Annual Reliability and Maintainability Symposium (RAMS)* (pp. 1-6). IEEE.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6769-6781).

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE transactions on big data*, *7*(3), 535-547.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in neural information processing systems*, *33*, 9459-9474.

Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and Bikel, D., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Nogueira, R. and Cho, K., 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

NVIDIA. NVIDIA Jetson Orin Nano 8GB Developer Kit. URL: https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/nano-super-developer-kit/

Facebook BlenderBot. URL: https://huggingface.co/facebook/blenderbot-400M-distill

Llama 3.2 3B. URL: https://huggingface.co/meta-llama/Llama-3.2-3B

Cross-encoder/ms-marco-TinyBERT-L2-v2. URL: https://huggingface.co/cross-encoder/ms-marco-TinyBERT-L2-v2

Sentence-transformers/all-MiniLM-L12-v2. URL: https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

BIOGRAPHIES

Lukmon Rasaq is a graduate research assistant and a Ph.D. candidate studying Industrial and Systems Engineering at North Carolina A&T State University. He earned his bachelor's degree in systems engineering from the University of Lagos, Nigeria, and his master's degree in computer science and information systems from the University of North Carolina Wilmington. His research focuses on implementing privacy and security techniques in smart manufacturing systems, human activity recognition, mitigating failures in commercial aircraft, addressing bias in AI systems, finetuning large language models (LLMs) for diverse domains, and developing and deploying RAG frameworks on edge devices for various use cases.

Madhuri Siddula is an Assistant Professor in the Department of Computer Science and Engineering at North Carolina Agricultural and Technical State University, USA. She earned her Ph.D. in Computer Science from Georgia State University in Atlanta, USA. She has authored or co-authored more than 30 research articles including IEEE IOT, WCMC,SCC and IEEE Access. Her research interests include Privacy Aware Computing, Cyber Security, IoT, LLMs and Social Networks

Om Prakash Yadav is a Professor and Chair in the Department of Industrial & Systems Engineering at North Carolina Agricultural and Technical State University, USA. He earned his Ph.D. in Industrial Engineering from Wayne State University in Detroit, USA. He has authored or co-authored more than 150 research articles in different areas, namely reliability, risk assessment, and design optimization. His research interests include reliability modeling and analysis, risk assessment, robust product/process design, optimization techniques, lean manufacturing, and Six Sigma methodologies. Om's research has been published in Reliability Engineering & System Safety, Quality, and Reliability Engineering International, Journal of Risk & Reliability, and so on.

Rhonda Walthall is a Senior Technical Fellow at Collins Aerospace Systems in Charlotte, NC, a division of RTX. In her role, she focuses on designing for PHM and Smart Products and supporting women in STEM roles. She is an industry-recognized leader in the development of industry standards and best practices for Integrated Aircraft Health Management (IAHM) solutions. She holds five PHM-related patents and numerous publications. She earned her BS degree in Aeronautical & Astronautical Engineering from Purdue University and her MBA from Pepperdine University. She is a member of the Purdue University AAE Industrial Advisory Council and was recognized as an Outstanding Aerospace Engineer in 2020. She is a Fellow of the PHM Society and a member of the Board of Directors. She is a Fellow of SAE International and a past member of the Board of Directors. She is a member of the Maintenance Programs Industry Group (MPIG), where she is advancing the use of IAHM as an alternative to scheduled aircraft maintenance. She is a member of Women in Aviation International, Society of Women Engineers, and Toastmasters International. Additionally, she has received numerous industry awards. Her book, *"Flight Paths to Success: Career Insights from Women Leaders in Aerospace",* won a prestigious award from the Independent Press.

Joseph Ensberg is a Technical Fellow of Applied Artificial Intelligence (AI) at RTX. In his role, he focuses on accelerating the rate at which RTX Business Units evaluate and adopt emerging AI capabilities to create discriminating products and services. Joseph holds numerous patents, trade secrets, and publications related to using advanced analytics and applied AI within aerospace and defense. He holds a Bachelor's degree in Mechanical and Aerospace Engineering from the University of California, Irvine (UCI), as well as a Master's and Ph.D. in Chemical Engineering from the California Institute of Technology (Caltech). In addition to applied AI, he is also an expert on prognostics and health management (PHM), environmental control system design, and atmospheric chemistry & physics.

Piyush Yadav is Chief Technologist for Applied AI in the Cross Cutting Technology Group at Collins Aerospace, Cork, Ireland, where he leads research on AI-powered edge cloud intelligence, generative AI, and data platform services for aerospace applications. He has over 15 years of experience spanning industry and academia, including research and teaching roles at the University of Galway and TCS Research, and holds a Ph.D. in computer science. Dr. Yadav has authored more than 60 peer-reviewed publications, including patents, journal articles, book chapters, and international conference papers, with research interests in generative AI, machine learning, IoT, cloud-edge AI, TinyML, complex event processing, and data analytics. His work focuses on developing real-world AI solutions that advance intelligent systems in aerospace and related domains.