





Approach Towards Semi-Automated Certification of Low Criticality ML-Enabled Airborne Applications

Chandrasekar Sridhar[†] , Vyakhya Gupta[†] , Prakhar Jain[†] , Karthik Vaidhyanathan[†] 
Software Engineering Research Center, IIT Hyderabad, India
 chandrasekar.s@research.iit.ac.in, vyakhya.gupta@students.iit.ac.in, prakhar.jain@research.iit.ac.in,
 karthik.vaidhyanathan@iit.ac.in

Abstract—As Machine Learning (ML) makes its way into aviation, ML-enabled systems—including low-criticality systems—require a reliable certification process to ensure safety and performance. Traditional standards, like DO-178C, which are used for critical software in aviation, don't fully cover the unique aspects of ML. This paper proposes a semi-automated certification approach, specifically for low-criticality ML systems, focusing on data and model validation, resilience assessment, and usability assurance while integrating manual and automated processes. Key aspects include structured classification to guide certification rigor on system attributes, an Assurance Profile that consolidates evaluation outcomes into a confidence measure the ML component, and methodologies for integrating human oversight into certification activities. Through a case study with a YOLOv8-based object detection system designed to classify military and civilian vehicles in real-time for reconnaissance and surveillance aircraft, we show how this approach supports the certification of ML systems in low-criticality airborne applications.

Index Terms—ML Certification, Low Criticality ML Certification

I. INTRODUCTION

The increasing integration of ML in aviation presents both opportunities and challenges, including applications with low criticality where flexibility and adaptability are valued but must still meet safety and reliability standards. Traditional certification standards such as DO-178C [1] were developed for deterministic rule-based software, emphasizing strong traceability from requirements to code, a task that is often challenging in practice. However, ML systems (MLS) operate differently, being inherently probabilistic and data-driven, meaning they learn from data rather than following predefined paths [2]. This fundamental shift has created gaps in existing certification standards [3], as they do not address the unique needs of ML, such as continuous model validation, data integrity, performance uncertainty handling [2]–[8].

The complex, dynamic nature of MLS introduces a major challenge in scaling certification processes. Manually certifying each component can be resource-intensive and prone to inconsistencies. Automation is therefore essential to improve efficiency and ensure reliable, repeatable assessments, which are crucial for ensuring the safety and reliability. Recent guidelines from the European Union Aviation Safety Agency (EASA) [9] highlight these gaps, recognizing that existing standards are insufficient for assessing ML behavior across diverse operational scenarios and addressing issues like data drift. These limitations hinder the deployment of ML technologies in airborne systems, where maintaining both safety and reliability is critical.

This paper proposes a semi-automated certification approach designed for low-criticality MLS in aviation. In this context, "low-

criticality" corresponds to DO-178C Level D, where software failures result in minor failure conditions for the aircraft. Our focus is to explore approaches for maintaining the validity of an MLS certification throughout its operational life. Unlike traditional software, MLS may require periodic re-evaluation to maintain certification standards. This paper makes the first step in addressing these requirements by proposing a certification process. We present a structured classification approach to tailor verification extensiveness, an *Assurance Profile* to quantify confidence, and methods for incorporating human oversight into certifying MLS in Level D criticality aviation. To effectively address these challenges, we make a few key assumptions, similar to those mentioned by in [10]. First, our focus is solely on the ML components of the system, while traditional, non-ML components are assumed to be certifiable under DO-178C without modification. Second, we do not address adaptive MLS that continue to learn and evolve while in operation, as these introduce further complexities. Our analysis is limited to systems where ML models are trained offline and remain static during operation.

To illustrate this approach, we examine a case study involving a YOLOv8-based object detection system [11], aimed at real-time classification in surveillance aircraft. The paper is structured as follows: Section II discusses related work and challenges in certifying ML for aviation. Section III describes the overview of the Air Sight case study. Section IV presents the semi-automated certification approach. Section V provides preliminary evaluation, and Section VI concludes with insights and future directions.

II. RELATED WORK

This work builds on previous research in MLS certification for low-criticality airborne applications, particularly Dmitriev et al. [10], who address MLS-specific certification challenges in aviation. Sculley et al. [12] further highlight how poor design decisions can increase maintenance costs and degrade performance, reinforcing the need for reliable design and testing practices. Tambon et al. [2] review certification strategies for safety-critical MLS, identifying gaps in addressing the probabilistic nature, continuous validation, and interpretability of ML, factors that we address with our semi-automated approach. Similarly, Pérez-Cerrolaza et al. [13] emphasize specialized certification methods for AI in safety-critical settings, highlighting the varied complexities of certifying MLS across sectors. [14]–[16] talk about testing and explaining deep models. For reliable MLS certification, we draw inspiration from the ML Test Score [17], structuring test cases and scoring to address challenges. This approach focuses on Level D criticality MLS certification, also adapting principles

[†] All the authors contributed equally to this work.

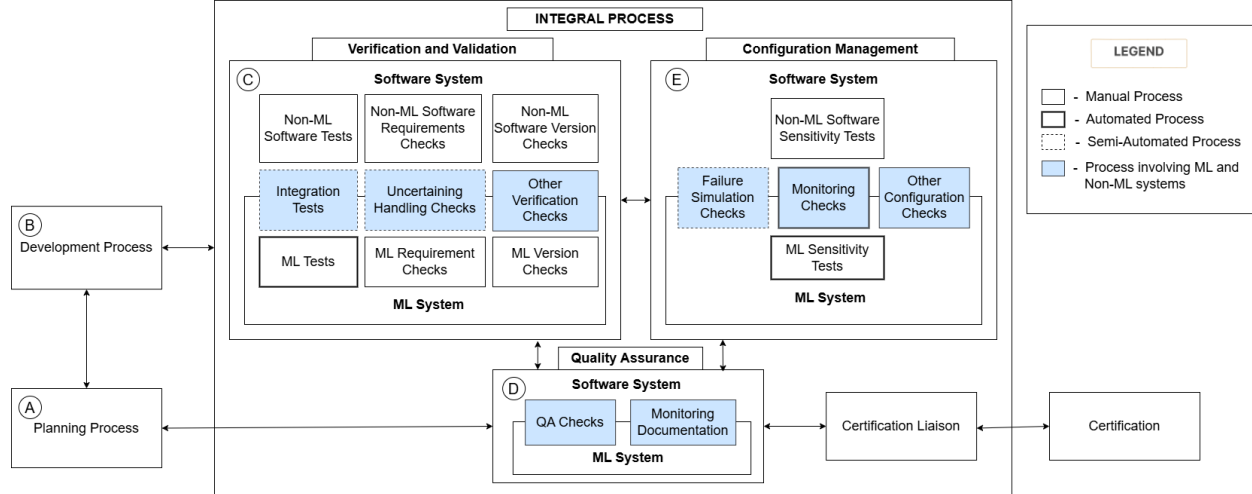


Fig. 1. Proposed Integrated Certification Process for ML and Non-ML System

from critical space systems [18] that emphasize traceability and validation for aviation needs. Additionally, Dmitriev et al. [19] explore the limitations of tool standards like DO-330 [20] and DO-200B [21] when applied to MLS, recommending adjustments to support MLS-specific workflows, including data management and model training. We also incorporate structured certification methods from Delseny et al. [22], which emphasize safety and reliability for MLS. This work also draws insights from recent research on best practices and standards proposed for ML-enabled systems and cyber-physical systems, which address challenges in ensuring robustness, systematic testing, and compliance throughout their lifecycle [23]–[29].

III. AIR SIGHT - CASE STUDY

This case study examines the certification process for a machine learning-based object detection system designed for low criticality airborne applications in a military context. The system, which utilizes a YOLOv8 model trained on a diverse dataset of military and civilian vehicles, aims to classify various types of air and ground assets in real-time. Intended for integration into reconnaissance and surveillance aircrafts, the system processes visual data to identify and classify vehicles of interest, enhancing situational awareness for pilots and operators. While not directly involved in flight-critical operations, the system’s outputs may inform tactical decisions, necessitating a robust certification process. This serves as a foundation for exploring how existing software certification standards, particularly DO-178C, can be extended to accommodate the unique characteristics of ML-based systems in low criticality airborne applications. The requirements for this certification approach were partially shaped by guidance from stakeholders specialized in the airborne system domain. Their expertise guided the exploration of extending existing software certification standards, particularly DO-178C, to accommodate the unique characteristics of ML-based systems.

IV. SEMI-AUTOMATED CERTIFICATION APPROACH

The semi-automated certification approach, as illustrated in Figure 1, addresses two primary challenges: *validating model con-*

sistency and assessing overall system resilience. This approach integrates both manual and automated processes across multiple stages, including Verification and Validation, Configuration Management, and Quality Assurance, which are exemplified through the Air Sight case study in the following sections.

A. Classification Approach for ML-enabled Systems

The approach uses a multi-axial classification which defines the certification requirements for MLS in aviation, tailoring rigor to system attributes. It is defined as $C = \langle c_{crit}, c_{aut}, c_{model} \rangle$ where: **a) System Criticality (c_{crit}):** follows DO-178C guidelines, ranging from Level A (highest criticality) to Level E (no impact on safety). This study focuses on Level D, representing low-criticality applications that still require basic safety and reliability standards. **b) Autonomy Level (c_{aut}):** assessed based on EASA’s levels [9], spanning from Level 1A (high human supervision) to Level 3B (full autonomy). Higher levels warrant stricter certification due to reduced human oversight. **c) ML Complexity (c_{model}):** Three levels adjust certification rigor based on model sophistication: **i) Level 1:** Simple models (e.g., linear or logistic regression) with high interpretability; requires minimal validation (V_1). **ii) Level 2:** Intermediate models (e.g., decision trees, random forests) with moderate adaptability; needs additional validation for reliability (V_2). **iii) Level 3:** Complex models (e.g., neural networks, CNNs) with nonlinear, high-dimensional decision boundaries; demands extensive validation due to challenges in interpretability (V_3).

Air Sight is classified as $C = \langle c_{crit} : D, c_{aut} : 2A, c_{model} : 3 \rangle$, meaning it operates at a low-criticality level, with moderate autonomy requiring human oversight, and utilizes a complex ML model like YOLOv8. This classification ensures the certification process is tailored to focus on appropriate levels of validation, robustness, and interpretability for safe and reliable operation.

B. Certification Layers

The proposed approach comprises of three layers, each addressing distinct assurance aspects for ML-enabled systems. These layers are based on the system’s classification, C . **i) Base Layer (DO-178C Processes):** covers requirements capture,

software development, configuration management, verification and integration. This layer establishes a solid foundation for certifying the non-ML components of the $c_{\text{crit}} = D$ system. **ii) ML Layer (ML-Specific Assurance):** focuses on ML-specific assurance, encompassing the entire ML lifecycle from dataset preparation and model training to hyperparameter optimization and performance evaluation. For $c_{\text{model}} = 3$, additional safeguards address increased non-linearity and dimensionality. **iii) Human Factors Layer (Human-AI Interaction Assurance):** targets the usability, transparency, and reliability of the AI components to support operator decision making. For $c_{\text{aut}} = 2A$ in Air Sight, we evaluate how the system enhances trust and effectiveness through usability testing in simulated environments, ensuring it improves situational awareness and reduces cognitive workload. The Base Layer follows standard practices, and the focus here is on the ML and Human Factors Layers, which apply specific methodologies to meet operational and safety requirements.

C. Certification Process

Automation is key to scalable, consistent evaluations in this certification approach. Solely manual certification is impractical for MLS due to model complexity and the vastness of data. Automating checks like data validation, performance monitoring, and robustness testing ensures reliable, ongoing validation. It also supports continuous compliance by detecting shifts in data relevance or model performance over time.

Central to our approach is the *Assurance Profile*, a transparent summary of compliance for the ML layer. It uses a structured scoring system to derive a final certification score, S_{cert} , based on manual reviews and automated checks. This score translates into a Confidence Level (σ) offering an intuitive assurance metric for stakeholders.

As illustrated in Figure 1, the approach includes key processes: Development (Dev), Verification and Validation (V&V), Quality Assurance (QA), and Configuration Management (CM). After each process, artifacts and evidence—such as Dataset Documentation, Model Configuration Reports, verification results, and monitoring logs—are generated to support the certification. Each process, denoted as $\text{Process}_i = \{\text{Dev}, \text{V\&V}, \text{QA}, \text{CM}\}$, is comprised various activities $\{\text{act}_1, \text{act}_2, \dots\}$, each evaluated through a series of manual and automated checks. The scores of manual evaluations, along with the percentage of automated checks passed, are aggregated to assess the performance each activity by producing S_{act_i} .

For each process, the score S_{process} is the weighted sum of its activities:

$$S_{\text{process}} = \sum_{i=1}^n w_{\text{act}_i} \cdot S_{\text{act}_i}$$

The weights for each activity, w_{act_i} , are determined as a factor of the system's classification C and contextual factors, including activity complexity and its impact on system performance, where S_{act_i} is the score of the i -th activity.

The final certification score S aggregates the scores from all processes, with each process weighted by w_{process_j} , which is similarly a factor of C and contextual influences:

$$S = \sum_{j=1}^4 w_{\text{process}_j} \cdot S_{\text{process}_j}$$

The certification confidence level $\sigma(S)$ categorizes the MLS score $S \in [0, 100]$ into five confidence levels:

$$\sigma(S) = \begin{cases} \text{Optimal Assurance} & : 90 \leq S \leq 100 \\ \text{Strong Assurance} & : 80 \leq S < 90 \\ \text{Moderate Assurance} & : 70 \leq S < 80 \\ \text{Limited Assurance} & : 60 \leq S < 70 \\ \text{Insufficient Assurance} & : S < 60 \end{cases}$$

Where: **Optimal Assurance** indicates maximum reliability; **Strong Assurance** suggests high performance with minor areas for improvement; **Moderate Assurance** meets criteria, with periodic review advised; **Limited Assurance** implies marginal reliability, requiring closer oversight; and **Insufficient Assurance** denotes significant deficiencies needing correction.

1) *Planning Process:* The planning process, as illustrated in Figure 1 (A), defines the foundational requirements for functional and non-functional aspects, forming the basis for traceability of the ML Layer. It establishes criteria for dataset diversity, label accuracy, and model performance, setting thresholds for key assurance checks. An uncertainty management strategy is included, setting confidence levels to handle ambiguous cases and guide human oversight.

For the Air Sight system, these requirements classify it as $C = \langle c_{\text{crit}} : D, c_{\text{aut}} : 2A, c_{\text{model}} : 3 \rangle$. This C , along with detailed requirements documentation, informs the weights assigned in the *Assurance Profile*. It also guides the core processes outlined in Figure 1, so that planning aligns with each layer's evaluation criteria.

2) *Development Process:* This process (refer Figure 1 (B)), focuses on building the ML dataset, model, and non-ML components, adhering to predefined requirements using automated tools for data integrity and model configuration. These tools verify dataset completeness, label accuracy, and class balance, while model architecture and hyperparameters are documented. Training processes must align with requirements. Unlike traditional software, MLS require thorough certification of the development activities for model reliability. This is less emphasized in lower-criticality systems but essential for robust validation. Integration documentation covers how ML and non-ML components interact, including interface specifications, data flow mappings, and system-level integration.

For Air Sight, the dataset is assessed for indicators like environmental diversity, which affects model adaptability. Some metadata lacks specific details on lighting or weather, lowering dataset score, but thorough preprocessing and augmentation like rotation and scaling improve the model's ability to detect objects under different perspectives. Integration documentation details how augmented data flows into the system's tactical display, ensuring real-time situational awareness without impacting operational safety. In the context of Level D criticality, it allows flexibility in dataset limitations, while Level 3 model complexity requires fine-tuning and detailed documentation of model adjustments.

3) *Verification and Validation:* This process, as shown in Figure 1 (C), is essential for building the *Assurance Profile*, blending traditional V&V processes with methods tailored for ML components. This process evaluates the ML Layer and ensures that it meets the required aviation safety and reliability standards. It includes the following key activities:

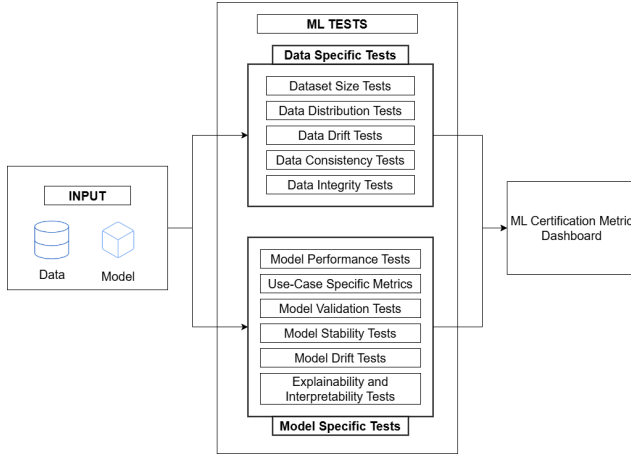


Fig. 2. Automated ML Testing

a) Automated Processes: Figure 2 shows our ML certification testing approach, covering key datasets and model checks for automated testing. Our ML certification approach includes automated testing to ensure reliability and robustness. *i) Model Performance* is evaluated using metrics like precision and F1-score across varied operational scenarios to validate real-time detection capabilities. *ii) Robustness Testing* examines the model’s performance under real-world edge cases such as occlusion, motion blur, and noise, ensuring functionality without adversarial scenarios. *iii) Dataset Certification* ensures dataset integrity, distribution, and anomaly detection, including data drift, to align with the operational environment and minimize biases. Automated checks assess statistical reliability, while human reviewers verify their outcomes and ensure proper documentation.

b) Semi-Automated Processes: System integration and uncertainty handling require human intervention for complex scenarios. Automated checks verify consistency, while human reviewers assess edge cases that automation may overlook. *i) System Integration* to verify seamless communication between ML and non-ML components across interfaces and data flows. *ii) Uncertainty Handling* assesses the system’s response to ambiguous cases, requiring human oversight for complex outputs and timely intervention. Due to the non-deterministic nature of AI, it is crucial to set acceptable error tolerances for expected scenarios. Additionally, fail-safe mechanisms must be in place to address unexpected or unsafe behavior, mitigating risks in line with aviation standards. *iii) Human Factors* testing incorporates usability assessments through simulated decision-making environments and operator feedback, ensuring the system aligns with user needs.

For Air Sight, key model metrics like precision, recall, mAP, accuracy were as per requirements, despite some class performance imbalance. Testing revealed degradation under conditions like noise and varying environmental factors. Dataset certification identified image property outliers, such as unusual lighting, which could affect model reliability. With $c_{\text{model}}=3$, extensive validation, $\langle V_3 \rangle$ was performed.

4) Quality Assurance: As depicted in Figure 1 (D), in a Level D MLS, QA ensures compliance with DO-178C standards across all certification layers, focusing on consistency

and stability. It integrates automated and human-supervised tools to verify adherence to the Base Layer, ML Layer, and Human Factors Layer, ensuring the reliability of both software components and ML models. Key QA activities include: *i) Post-Certification Operational Monitoring:* QA ensures dataset relevance, model integrity, and ML accuracy through testing for robustness and interpretability throughout operational use. A structured monitoring plan tracks performance and user feedback, with recertification triggers for significant updates to maintain compliance. Automated monitoring detects data shifts, but human review determines their significance and corrective actions. *ii) Adherence Audits & Documentation Reviews:* Routine audits confirm that development, testing, and human factors assessments meet standards, covering ML training, dataset management, and interface design. Formal reviews ensure proper documentation of model versions, datasets, and changes for traceability and compliance. *iii) Usability Assessment:* Evaluations ensure human-AI interactions enhance situational awareness and reduce cognitive load, supporting effective decision-making.

In the Air Sight system, QA activities identified key areas for improvement, particularly in post-certification monitoring and usability. Low usability scores resulted from difficulties in interpreting AI decisions in complex scenarios. Monitoring also revealed label drift when encountering previously unrepresented vehicles or conditions. Despite these challenges, the QA framework maintains compliance and provides insights for improving model performance and user interaction.

5) Configuration Management: CM process (refer Figure 1 (E)) for Level D MLS ensures consistency across all components, including the user interface, ML model, and datasets. By combining automated and human-supervised activities, the CM process focuses on managing significant updates that impact performance and human interaction. Key CM activities include:

i) Configuration Identification and Version Control: All components, including ML models, datasets, and human-interface elements, are identified and documented to create a stable baseline. Major updates, such as model retraining or interface adjustments, undergo appropriate version control. *ii) Configuration Audits and Compliance:* Regular audits verify that all changes align with system standards and do not negatively affect the user experience or system performance. These checks prevent updates from disrupting operator situational awareness or increasing the complexity of human-AI interactions. *iii) Post-Certification Update Management:* Procedures exist for updates to both the ML model and user interface after certification. Significant updates follow steps for documentation, review, and a potential recertification.

For Air Sight, these processes are essential for managing updates to the model, datasets, and user interface components. While major changes to YOLOv8 and the dataset are properly version-controlled, the current process lacks a structured approach for post-certification updates due to the experimental nature of the system. This gap in update management impacts the overall score in this domain.

V. PRELIMINARY EVALUATION

The YOLOv8 model, used for object detection in the Air Sight case study, was fine-tuned on a Nvidia GTX 1080 Ti with 12 GB VRAM using the [30] dataset, ensuring that it could accurately classify various types of air and ground assets. Our approach to

certify the system was partially realized using Deepchecks [31] on a Ubuntu 22.04 machine with 16GB RAM powered by 4.7 GHz 12 Gen Intel i7 processor. Deepchecks was used specifically to assess various types of drifts. Table I offers a comprehensive summary of the Air Sight system’s certification journey, synthesizing scores from automated, semi-automated, and manual reviews across critical assurance dimensions. These scores were determined through structured assessments, combining automated tools with review sessions, where the authors collaborated with stakeholders to ensure alignment with certification criteria. Through our semi-automated certification approach¹, the example below illustrates the Air Sight system’s current status against the outlined criteria for a Level D classification.

TABLE I
ASSURANCE PROFILE FOR AIR SIGHT SYSTEM

Air Sight System Certification Summary (DO-178C Level D)			
Process	Score (100)	Weight	Weighted Score
Dataset Quality	80.0	0.40	32.0
Model Documentation	82.0	0.35	28.7
Integration Documentation	70.0	0.25	17.5
Total Development Score			78.2
2. Verification & Validation (V&V)			
Model Performance	95.0	0.25	23.8
Robustness Testing	92.0	0.25	23.0
Dataset Certification	84.8	0.20	17.0
System Integration	88.0	0.15	13.2
Human Factors	87.0	0.15	13.0
Total V&V Score			90.0
3. Quality Assurance (QA)			
Post-Certification Monitoring	55.0	0.35	19.3
Usability Assessment	52.0	0.35	18.2
Audits and Reviews	50.0	0.30	15.0
Total QA Score			52.5
4. Configuration Management (CM)			
Version Control	65.0	0.40	26.0
Configuration Identification	60.0	0.35	21.0
Change Management	58.0	0.25	14.5
Total SCM Score			61.5
Final Certification Summary			
Development	78.2	0.30	23.5
V&V	90.0	0.35	31.5
QA	52.5	0.20	10.5
SCM	61.5	0.15	9.2
Final Certification Score			74.7

A. Certification Details

The *Assurance Profile* for the Air Sight system presents a detailed evaluation of its certification readiness, resulting in a Final Assurance Score of 74.7%, indicating a Moderate Assurance level as per the scoring benchmark IV-C. Strong results in the V&V and Development stages reflect the system’s robust dataset quality, clear model documentation, and effective validation processes. However, the lower scores in QA and CM suggest areas that need improvement, particularly in post-deployment monitoring and version control.

This initial certification assessment demonstrates that Air Sight meets the compliance criteria for DO-178C Level D criticality, suitable for applications with lower safety risks. Given its classification *C*, full recertification is not necessary within a typical operational lifecycle. Instead, Air Sight will undergo targeted drift checks, which are periodic evaluations to ensure the model’s performance remains consistent over time, especially in

¹<https://github.com/sa4s-serc/MLCert>

response to changes in data patterns or operational environments. This avoids the burden of a full certification process while maintaining a focus on detecting significant shifts early.

The need for improved QA signals the importance of strengthening ongoing audit processes to ensure sustained reliability. With the system’s autonomy and ML complexity levels, these- coupled with frequent drift checks— will help proactively manage any changes, ensuring that Air Sight continues to meet its operational requirements effectively and safely.

B. Recertification Triggers

To ensure the continued safety and reliability of the Air Sight system, effective recertification planning is crucial. The conditions to prompt this, as outlined in CM and QA in Section IV-C, are: *i) Performance Degradation*: Recertification is required if operational accuracy drops below desired threshold. *ii) Dataset Shift*: A significant change (over 30%) in dataset distribution or model architecture triggers recertification to evaluate data drift or model evolution. *iii) Environmental Changes*: Major updates to operational conditions, such as new deployment scenarios or mission requirements. *iv) Updates to Uncertainty Handling*: Changes to uncertainty management strategies” revised confidence thresholds or new failure modes.

C. Evaluation of our Certification Approach

To assess the validity of our approach, we introduced gaussian noise into the validation set derived from the [30] dataset. This simulated real-world conditions where sensor or environmental inconsistencies may degrade data quality. Despite these perturbations, the MLS achieved the performance metrics: 79% precision and 75% recall and a mAP score of 81%. Compared to the unperturbed dataset, the performance degradation was minimal, reinforcing the MLS’s robustness under real-world conditions. While prediction drift was observed during testing, it remained within the acceptable threshold for Level D criticality systems (<30%), negating the need for recertification. The evaluation confirmed that the ML component of the system maintains acceptable performance and compliance with certification criteria under perturbed conditions, validating its readiness for deployment in Level D applications.

VI. CONCLUSION

The integration of ML-based systems into aviation introduces new certification challenges, as illustrated by the Air Sight case study. Traditional standards like DO-178C lack the adaptability needed for certifying data-driven, evolving ML systems (MLS). Our semi-automated certification approach addresses these gaps by focusing on continuous model performance evaluation, data quality management, and resilience in varied operational contexts. We also introduce an *Assurance Profile* for MLS, akin to a nutrition label. This *Assurance Profile* provides a structured way forward in certifying MLS, capturing the nuanced characteristics of MLS.

We plan to extend our approach to accommodate a broader range of use cases, enabling its application to more complex MLS across various domains and criticality levels. Additionally, we aim to explore different values for weights to assess their effectiveness in contributing to system safety. These efforts will help our semi-automated certification approach and enhance its applicability across MLS with varying safety and reliability requirements.

REFERENCES

- [1] Radio Technical Commission for Aeronautics (RTCA), "DO-178C: Software Considerations in Airborne Systems and Equipment Certification," 2011. RTCA DO-178C.
- [2] F. Tambon, G. Laberge, L. An, A. Nikanjam, P. S. N. Mindom, Y. Pequignot, F. Khomh, G. Antoniol, E. Merlo, and F. Laviolette, "How to certify machine learning based safety-critical systems? a systematic literature review," *Automated Software Engineering*, vol. 29, no. 2, p. 38, 2022.
- [3] E. Jenn, A. Albore, F. Mamalet, G. Flandin, C. Gabreau, H. Delseny, A. Gauffriau, H. Bonnin, L. Alecu, J. Pirard, *et al.*, "Identifying challenges to the certification of machine learning for safety critical systems," in *European congress on embedded real time systems (ERTS 2020)*, 2020.
- [4] M. Staples, L. Zhu, and J. Grundy, "Continuous validation for data analytics systems," in *Proceedings of the 38th International Conference on Software Engineering Companion, ICSE '16*, (New York, NY, USA), p. 769–772, Association for Computing Machinery, 2016.
- [5] Z. Chain and D. Alexander, "Data integrity challenges and solutions in machine learning- driven clinical trials," 08 2023.
- [6] E. Caveness, P. S. GC, Z. Peng, N. Polyzotis, S. Roy, and M. Zinkevich, "Tensorflow data validation: Data analysis and validation in continuous ml pipelines," in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 2793–2796, 2020.
- [7] J. Bosch, H. H. Olsson, and I. Crnkovic, "Engineering ai systems: A research agenda," *Artificial intelligence paradigms for smart cyber-physical systems*, pp. 1–19, 2021.
- [8] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.
- [9] EASA, "Easa artificial intelligence (ai) concept paper issue 2: Guidance for level 1&2 machine learning applications," March 2024.
- [10] K. Dmitriev, J. Schumann, and F. Holzapfel, "Toward certification of machine-learning systems for low criticality airborne applications," in *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, pp. 1–7, IEEE, 2021.
- [11] G. Jocher *et al.*, "Yolov8." <https://github.com/ultralytics/ultralytics>, 2023.
- [12] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15*, (Cambridge, MA, USA), p. 2503–2511, MIT Press, 2015.
- [13] J. Perez-Cerrolaza, J. Abella, M. Borg, C. Donzella, J. Cerquides, F. J. Cazorla, C. Englund, M. Tauber, G. Nikolakopoulos, and J. L. Flores, "Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–40, 2024.
- [14] Y. Yang and M. Rinard, "Correctness verification of neural networks," 2022.
- [15] K. Pei, Y. Cao, J. Yang, and S. Jana, "Deepxplore: Automated whitebox testing of deep learning systems," in *Proceedings of the 26th Symposium on Operating Systems Principles, SOSP '17*, p. 1–18, ACM, Oct. 2017.
- [16] Y. Tian, K. Pei, S. Jana, and B. Ray, "Deeptest: Automated testing of deep-neural-network-driven autonomous cars," 2018.
- [17] E. Breck, S. Cai, E. Nielsen, M. Salib, and D. Sculley, "The ml test score: A rubric for ml production readiness and technical debt reduction," in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 1123–1132, 2017.
- [18] A. Petrucci, F. Basciani, and P. Pelliccione, "Ai/ml for safety-critical software: the case of the space domain," *IEEE Software*, pp. 1–8, 2024.
- [19] K. Dmitriev, F. Kaakai, M. Ibrahim, U. Durak, B. Potter, and F. Holzapfel, "Tool qualification aspects in ml-based airborne systems development," in *Software Engineering 2023 Workshops*, pp. 208–221, Bonn: Gesellschaft für Informatik e.V., 2023.
- [20] Radio Technical Commission for Aeronautics (RTCA), "DO-330: Software Tool Qualification Considerations," 2011. RTCA DO-330.
- [21] Radio Technical Commission for Aeronautics (RTCA), "DO-200B: Standards for Processing Aeronautical Data," 2015. RTCA DO-200B.
- [22] H. Delseny, C. Gabreau, A. Gauffriau, B. Beaudouin, L. Ponsolle, L. Alecu, H. Bonnin, B. Beltran, D. Duchel, J.-B. Ginestet, *et al.*, "White paper machine learning in certified systems," *arXiv preprint arXiv:2103.10529*, 2021.
- [23] H. Bello, D. Geißler, L. Ray, S. Müller-Divéky, P. Müller, S. Kittrell, M. Liu, B. Zhou, and P. Lukowicz, "Towards certifiable ai in aviation: landscape, challenges, and opportunities," *arXiv preprint arXiv:2409.08666*, 2024.
- [24] M. Gariel, B. Shimanuki, R. Timpe, and E. Wilson, "Framework for certification of ai-based systems," *arXiv preprint arXiv:2302.11049*, 2023.
- [25] L. Rierison, *Developing safety-critical software: a practical guide for aviation software and DO-178C compliance*. CRC Press, 2017.
- [26] M. Sivakumar, A. B. Belle, K. K. Shahandashti, O. Odu, H. Hemmati, S. Kpodjedo, S. Wang, and O. O. Adesina, "I came, i saw, i certified: some perspectives on the safety assurance of cyber-physical systems," *arXiv preprint arXiv:2401.16633*, 2024.
- [27] J. Chandrasekaran, T. Cody, N. McCarthy, E. Lanus, and L. Freeman, "Test & evaluation best practices for machine learning-enabled systems," *arXiv preprint arXiv:2310.06800*, 2023.
- [28] M. Anisetti, C. A. Ardagna, N. Bena, and E. Damiani, "Rethinking certification for trustworthy machine learning-based applications," *IEEE Internet Computing*, 2023.
- [29] R. Sabatini, E. Blasch, I. Majid, A. Gardi, and A. Roy, "Application and certification challenges for ai/ml techniques in safety critical avionics systems," 09 2022.
- [30] M. V. O. Detection, "Military vehicles object detection dataset." <https://universe.roboflow.com/military-vehicle-object-detection/military-vehicles-object-detection>, may 2023. visited on 2024-11-06.
- [31] S. Chorev, P. Tannor, D. B. Israel, N. Bressler, I. Gabbay, N. Hutnik, J. Liberman, M. Perlmutter, Y. Romanyshyn, and L. Rokach, "Deepchecks: A library for testing and validating machine learning models and data," *Journal of Machine Learning Research*, vol. 23, no. 285, pp. 1–6, 2022.