AI-Powered Classification of Training-Related Aviation Accidents: Performance Evaluation of a Gemini-Based TSB Report Filtering System

Leyao Cui, Jacob Nowak, Shi Cao.
Waterloo Institute for Sustainable Aeronautics (WISA), University of Waterloo
Department of Systems Design Engineering, University of Waterloo

Abstract:

Aviation accident reports are an important source of information for identifying safety issues and preventing future occurrences. However, finding reports related to contributing factors can be time-consuming, since investigation databases do not always include detailed labels for every safety issue discussed in the full report. This study evaluates an artificial intelligence (AI)-based classification system designed to identify General Aviation (GA) accident reports that contain potential training-related systemic deficiencies.

The system analyzed 1,312 aviation accident investigation reports published by the Transportation Safety Board of Canada (TSB) between 1991 and November 29, 2025, using Google's Gemini Flash API as part of a two-stage filtering process.

Stage 1 filtered for GA fixed-wing aircraft accidents (including Air Taxi operations), reducing the dataset from 1,312 to 595 reports (45.4%). Stage 2 applied a rigorous 4-filter classification protocol to identify training-related systemic gaps, ultimately identifying 97 preventable training-related accidents (7.4% of original dataset).

Manual validation on a stratified sample of 178 reports (30% of Stage 2 outputs) demonstrated strong performance: 88.8% overall accuracy, 75.0% precision, and 96.4% recall. The high recall rate ensures comprehensive capture of genuine training gaps with minimal false negatives, supporting the research objective of identifying preventable accidents through targeted training interventions.

Keywords: aviation accident analysis; aviation safety; manual validation; flight training; training deficiencies; Large Language Models (LLMs); accident report classification.

APA style citation as: Cui, L., Nowak, J., Cao, S. (2026). AI-Powered Classification of Training-Related Aviation Accidents: Performance Evaluation of a Gemini-Based TSB Report Filtering System. WISA Technical Report (2026-001). Waterloo Institute for Sustainable Aeronautics, University of Waterloo.

1. Introduction:

Aviation accident investigation plays an important role in improving the safety of the air transportation system. Although aviation accidents are relatively uncommon, their consequences can be severe, making it important to understand the operational, human, and organizational

factors that contribute to them. Accident reports provide detailed records of these factors, allowing researchers to learn from past occurrences and reduce the likelihood in the future.

In Canada, aviation accidents and incidents are investigated by the Transportation Safety Board of Canada (TSB), an independent federal agency responsible for advancing transportation safety (Transportation Safety Board of Canada, 2020). During an investigation, the TSB collects and analyzes evidence to determine what happened, identify the causes and contributing factors, document safety deficiencies, and share its findings with the public. The completed reports are published in the TSB's online database, where they are available to industry stakeholders, researchers, pilots, and the public (Transportation Safety Board of Canada, 2020).

However, the usefulness of these reports depends on how easy relevant information can be found. For example, researchers and investigators might want to find groups of accidents involving specific training deficiencies or procedural issues. While the reports contain detailed information, the database does not include complete labels for every contributing factor discussed. As a result, researchers must look through titles, occurrence details, keyword searches, and manually review individual reports. With both the number and length of investigation reports growing, searching a list of relevant reports can take much time and effort.

Artificial intelligence (AI) offers a potential way to reduce this workload by reviewing investigation reports and classifying them based on their contributing factors. A systematic review by Nanyonga et al. (2025) explores the application of machine learning techniques in aviation safety, including post-accident analysis, predictive safety applications, and incident detection. Their findings suggest that machine learning could be used to support large data set analysis and help find critical recurring patterns.

More specifically, Ziakkas and Pechlivanis (2023) examined how artificial intelligence could be interpreted by comparing AI-generated analyses with those produced by human experts. Their findings showed that while AI can help to classify accidents, human oversight remains important, especially when dealing with complex accident causes. Similarly, Dillon et al. (2024) applied AI to large amounts of information from incident-reporting systems to see how safety-related information can be extracted and support human learning.

Recent research has also explored how AI can support the accident investigation process itself. For example, Nakanishi and Ono (2025) investigated the use of AI to model the tacit knowledge used by experienced air accident investigators, with the goal of improving investigative decision-making. Building on this work, Nakanishi et al. (2026) developed an analytical framework based on large language models (LLMs) to identify accident factors and their relationships from aircraft accident investigation records. Their findings suggest that generative AI could help structure and review causal relationships before the information is reviewed by human investigators.

The goal of the current study is to further demonstrate that AI has the potential to help researchers analyze large amounts of complex aviation safety information. This paper will evaluate the performance of an AI-powered classification system designed to identify General Aviation (GA) accident reports where training-related deficiencies may have contributed to the

occurrence. The study analyzed 1,312 TSB aviation accident investigation reports published between 1991 and November 29, 2025. The objective was to evaluate whether an AI-based classification approach could reliably identify a focused subset of reports containing potential systemic training gaps while reducing the manual effort required to search the broader database. By enabling more efficient identification of recurring training-related deficiencies, the system may support future research in pilot education, training design, and aviation risk mitigation.

2. Methodology

2.1 Data Source and Filtering Pipeline

Data Source:

The source dataset comprised all aviation accident reports published on the Transportation Safety Board of Canada (TSB) website from 1991 onwards, totaling 1,312 investigation reports. This represents the complete publicly available Canadian aviation accident database for the study period, including all aircraft types (fixed-wing, rotary-wing, gliders, balloons, UAVs) and all operation categories (commercial airlines, general aviation, military, etc.).

Table 1. Two-stage filtering pipeline and report counts retained at each stage.

Stage	Objective	Method	Reports	% Retained
Input Dataset	All TSB Aviation Reports (1991+)	—	1,312	100%
Stage 1 Filter	Identify GA Fixed-Wing Operations	AI classification of aircraft type and operation category	595 (591 analyzed)	45.4%
Exclusions	Removed: Commercial Airlines, Helicopters, Gliders, UAVs, Military	—	717	—
Stage 2 Analysis	Identify Training-Related Systemic Gaps	4-filter causal analysis protocol	97	7.4% of original
Exclusions	Removed: Pilot error, mechanical failures, weather, non-causal findings, no full report error	—	498	—

Note on Air Taxi Inclusion:

While Air Taxi operations are technically classified as commercial aviation, they were included alongside pure General Aviation due to shared operational characteristics: similar aircraft types (Cessna 185, DHC-2 Beaver, Piper Navajo), common training pipelines, and analogous training challenges. Air Taxi operations constitute 46.9% (277 accessible reports, 278 total) of the Stage 1 filtered dataset. Detailed rationale is provided in **Appendix 10.1.**

Note on Data Completeness:

Of the 595 reports that passed Stage 1 filtering, four reports (0.7%) were inaccessible due to lack of full report links on the TSB website (SSA9301, SA9401, SIIA0501, A11Q0136). These were automatically classified as is_systemic_gap=False with Gap_Type='No Full Report'. The remaining 591 reports (99.3%) were successfully analyzed by the AI system.

2.2 Stage 1 Classification Prompt: Scope Filtering

Stage 1 employed a straightforward classification prompt to extract aircraft category and operation type from each TSB report.

Classification Criteria:

- Aircraft Category: Fixed-Wing (Airplane) vs. Rotary-Wing (Helicopter/Gyroplane) vs. Other (Glider/Balloon)
- Operation Type: General Aviation (Private, Training, Aerial Work, Air Taxi) vs. Commercial Airline (Scheduled 705, Commuter 704) vs. State/Military

Retention Logic:

Reports were retained (Is_GA_FixedWing = True) if they met both conditions:

(1) Aircraft Category = Fixed-Wing AND

(2) Operation Category = General Aviation OR Air Taxi.

This logic excluded scheduled airlines, helicopters, gliders, military operations, and ground vehicle incidents.

Technical Implementation: The AI extracted structured JSON output including ReportID, Aircraft_Category, Aircraft_Type (e.g., 'Cessna 172'), Operation_Category (e.g., ‘Airline/Commuter’), Operation_Detail (e.g., 'Flight Training'), and the boolean Is_GA_FixedWing flag. Only reports flagged as True proceeded to Stage 2.

2.3 Stage 2 Classification Prompt: Training Gap Analysis

Stage 2 employed a comprehensive 4-filter classification protocol designed to identify genuine systemic training deficiencies while minimizing false positives from pilot error, mechanical failures, and incidental safety findings.

TSB Report Structure Awareness:

The AI prompt incorporates understanding of TSB report structure, which distinguishes between:

- "Findings as to causes and contributing factors": Direct reasons the accident occurred
- "Findings as to risk": Latent unsafe conditions that did not necessarily cause this specific accident

Critical Rule: A systemic gap is only classified as TRUE if it appears as a CAUSE or contributing factor, not merely a RISK. This structural awareness prevents false positives from incidental safety recommendations.

Four-Filter Decision Protocol:

Each report must pass all four filters sequentially. Failure at any filter results in classification as is_systemic_gap = False.

Filter 1 - Responsibility Check (Individual vs. System):

Distinguishes whether the pilot made a mistake because they ignored their training (pilot error) or because the training system never taught them the required skill (systemic gap). If the pilot should have known better based on standard licensing requirements, the report fails this filter.

Filter 2 - Explicit Evidence Filter ('No Inference' Rule):

Requires explicit textual evidence that training or procedures were missing. Merely describing a hazard (e.g., 'black hole illusion') without stating 'the syllabus did not include this' results in failure. Accepted phrases include 'no procedure,' 'did not include,' 'not in syllabus,' 'not required by regulation.'

Filter 3 - Solvability Filter ('Hardware vs. Courseware' Test):

Distinguishes trainable gaps from untrainable hardware/regulatory deficiencies. Missing placards, warning labels, GPWS equipment, performance charts, or regulatory requirements are hardware/law gaps that cannot be solved through pilot training. Only missing procedures, instructions, checklist steps, or syllabus modules pass this filter.

Filter 4 - Risk-to-Cause Bridge (Causality Check):

Ensures identified gaps are causally linked to the accident. If a gap appears in 'Findings as to Risk' but the 'Findings as to Causes' attributes the accident to unrelated pilot error (e.g., ignored warning horn), the gap fails unless the Analysis section explicitly bridges the gap to the pilot's error (e.g., 'The lack of manual guidance explains why the pilot did not increase power').

Few-Shot Calibration Examples:
The prompt includes three concrete examples to calibrate AI decision-making:

- Example 1 (False): Aircraft not required to have a placard → Hardware/regulatory gap, not training
- Example 2 (False): Checklist missing speed brake item (Risk) + Crew ignored gear warning (Cause) with no causal bridge → Incidental finding
- Example 3 (True): Missing POH procedure (Risk) + Pilot did not increase power (Cause) + Analysis states 'lack of manual guidance explains the failure' → Causal systemic gap

Additional classification criteria details and concrete case examples from the TSB dataset are provided in **Appendix 10.2**.

2.4 Manual Validation Protocol

To evaluate AI classification accuracy, we manually reviewed a subset of 591 accessible reports from Stage 2 outputs. The validation sample was designed using stratified random sampling to ensure representative coverage.

Stratification Approach:
Reports were grouped into strata based on two dimensions:

- AI Classification: Training-Related (is_systemic_gap = True) vs. Non-Training (is_systemic_gap = False)
- Confidence Score: The AI assigns each classification a confidence score from 1-10, with 10 being most confident

This creates strata such as: 'Training-Related with Confidence 10,' 'Training-Related with Confidence 9,' 'Non-Training with Confidence 10,' etc.

Sampling Rules:

- Small Strata (<50 reports): Validate 100% of reports to ensure no rare patterns are missed
- Large Strata (≥50 reports): Randomly select exactly 50 reports for validation

Statistical Validity:

This stratified sampling approach provides statistically sound performance estimates for several reasons:

- Complete Coverage of Small Strata: Validating 100% of strata with <50 reports ensures no rare classification patterns are missed
- Adequate Sample Size for Large Strata: Random samples of 50 reports provide statistically reliable estimates (at 95% confidence level, margin of error ≈±14% for binary classification)
- Balanced Representation: Higher validation rate for Training-Related cases (74%) vs. Non-Training (21%) appropriately prioritizes verification of potential training gaps, which have greater research consequence

- Elimination of Selection Bias: Random sampling within each stratum ensures each report has equal probability of selection, preventing systematic bias
- Conservative Estimates: Stratified sampling typically reduces sampling error compared to simple random sampling, meaning the reported performance metrics likely represent conservative lower-bound estimates of true model performance

Table 2. Validation sample breakdown by AI classification and confidence score.

AI Classification	Confidence Score	Total in Stratum	Sample Size	Sampling Method
Training-Related	10	17	17	100% (small stratum)
Training-Related	9	75	50	Random sample
Training-Related	8	5	5	100% (small stratum)
Training Subtotal	—	97	72	74% overall
Non-Training	10	281	50	Random sample
Non-Training	9	207	50	Random sample
Non-Training	8	5	5	100% (small stratum)
Non-Training	3	1	1	100% (small stratum)
Non-Training Subtotal	—	494	106	21% overall
Total Validated	—	591	178	30% of dataset

2.5 Output Data Structure

Beyond binary classification, the AI system generates structured metadata for each report, providing actionable insights for training intervention design.

Key Output Fields:

- Gap_Type: Categorizes the deficiency (Missing Syllabus / Manual-SOP Omission / Fidelity Issue / N/A)

- Specific_Missing_Training: Detailed description of the identified gap (e.g., 'Type-specific multi-engine asymmetric power management after feathering')
- Confidence_Score: AI's confidence in the classification (1-10 scale, where 10 indicates highest certainty)
- Suggested_Intervention: Proposed remedial training or procedural change (e.g., 'Mandate rigorous simulation of engine failure requiring immediate maximum power application')
- Evidence_Quote: Direct quote from TSB report supporting the classification, enabling verification and citation
- Metadata: Aircraft_Type, Operation_Detail, Year, Link (direct URL to full TSB report)

3. Results

3.1 Performance Metrics

Manual validation of 178 reports (30% of Stage 2 outputs) yielded the following confusion matrix and derived performance metrics:

Table 3. Confusion matrix and derived performance metrics for the 178 validated reports.

Metric	Count	Formula	Value
True Positives (TP)	54	AI=True ∩ Manual=Match	—
False Positives (FP)	18	AI=True ∩ Manual=Mismatch	—
True Negatives (TN)	104	AI=False ∩ Manual=Match	—
False Negatives (FN)	2	AI=False ∩ Manual=Mismatch	—
Total	**178**	**—**	
Precision	—	TP / (TP + FP)	75%
Recall (Sensitivity)	—	TP / (TP + FN)	96.4%
Overall Accuracy	—	(TP + TN) / Total	88.8%

Interpretation:

Accuracy (88.8%):

The model correctly classified 158 of 178 validated reports, demonstrating strong overall performance across both training-related and non-training categories.

Precision (75%):

Of 72 reports flagged as training-related by the AI, 54 were confirmed upon manual review. This indicates a three-in-four probability that AI-identified gaps represent genuine systemic deficiencies.

Recall (96.4%):

The model successfully identified 54 of 56 validated training-related accidents. High recall is critical for the research objective, ensuring comprehensive capture of preventable training gaps with minimal false negatives (only 2 genuine cases missed).

Error Analysis:

False Positives (18 cases):

These represent borderline situations where training deficiencies were mentioned in the report but were not the primary causal factor, or where TSB language was ambiguous about whether a systemic gap existed versus individual pilot proficiency issues.

False Negatives (2 cases):

Only 2 genuine training-related cases were missed, likely representing subtle gaps described using non-standard terminology that did not trigger the AI's explicit evidence filter, or implicit deficiencies requiring deeper inference than the prompt permits.

4. Performance Analysis

4.1 Precision-Recall Relationship

The model demonstrates 75.0% precision and 96.4% recall, representing different aspects of classification performance.

Precision (75.0%) indicates that three-quarters of AI-flagged training gaps are confirmed upon manual review, with the remaining quarter representing borderline cases requiring expert adjudication.

Recall (96.4%) indicates that the model successfully identifies the vast majority of genuine training-related accidents, with only 2 cases missed out of 56 validated training-related reports.

This precision-recall relationship means that the model casts a wide net (high recall) while accepting some false positives (moderate precision). The operational implication is that AI-flagged cases should undergo human review before final determination.

4.2 Confidence Score Correlation

Analysis of validation results reveals strong correlation between AI confidence and classification accuracy (see **Appendix 10.3** for detailed breakdown). Reports assigned confidence scores of 10 achieved 98.5% validation accuracy (66 of 67 matches), while confidence 9 reports showed 84.0% accuracy (84 of 100 matches). This pattern validates the use of confidence scores for prioritizing human review, with high-confidence classifications (≥9) demonstrating reliable performance and lower-confidence classifications (<8) requiring mandatory expert verification.

The majority of true positive training-related cases received confidence scores of 9-10, indicating the model's pattern recognition is particularly reliable for clear-cut systemic gaps with explicit textual evidence and causal linkage in the TSB Analysis sections.

5. Comparative Analysis

5.1 AI Model vs. TSB Website Filtering

Table 4. Comparison of the AI model and the TSB website search function.

Capability	AI Model	TSB Website
Search Interface	Semantic multi-filter analysis	Keyword-based text search (single or multiple keywords)
Keyword Search Results	Not applicable (semantic analysis)	'training': 6 results (3 TP) 'train': 7 results (3 TP) ‘procedure’: 0 results ‘syllabus’: 0 results ‘simulator’: 0 results
Total Training-Related Identified	97 reports	~6 reports (highly keyword-dependent)
Detection Rate	7.4% of database	~0.5% of database
Detection Improvement	—	16× fewer cases detected
Semantic Understanding	Context-aware causal analysis	None (exact keyword matching only)
Processing Time (1,312 reports)	~5 minutes automated	Instant filtering
Consistency	Deterministic algorithm	Consistent keyword matching
Scalability	Parallel processing	Instant for filtering, linear for manual review

Critical Finding: The TSB website's keyword search dramatically under-detects training-related accidents because it only searches the brief summary text displayed under each report title, not the full investigation text. This limitation explains why searches for training-related terminology return minimal results: "training" returns 6 results, "train" returns 7 results, while "syllabus", "procedure", "manual", and "instruction" each return 0 results.

Most training gap discussions appear in the detailed "Analysis" and "Findings" sections of TSB reports, which are not indexed by the website search function. In contrast, the AI model processes the complete report text, enabling identification of training-related language regardless of location within the document. This structural difference accounts for the 16-fold detection improvement (97 AI-identified cases vs. ~6 keyword results), as the AI can recognize training

gaps described using varied terminology ("lack of procedure," "not required by syllabus," "inadequate preparation," "simulator did not replicate") buried deep within report analysis sections.

5.2 AI Model vs. Manual Expert Review

Table 5. Comparison of the AI model and manual expert review.

Criterion	AI Model	Manual Expert Review
Processing Time (Full Dataset)	~10 minutes total (Stage 1: ~5 min, Stage 2: ~5 min)	80-120 hours (1,312 reports × 4-6 min/report)
Consistency	High (deterministic algorithm)	Variable (reviewer-dependent)
Scalability	Excellent (parallel processing)	Poor (linear with reports)
Cost (Full Dataset)	$40-80 (API costs)	$4,000-6,000 (expert labor @ $50/hr)
Accuracy	88.8% (validated on sample)	100% (by definition)
Structured Metadata	Automatic extraction\	Note taking by reviewers
Fatigue Factor	None (consistent performance)	Significant (quality degrades over long sessions)

The AI model achieves ~99% time reduction compared to manual review (10 minutes vs. 80-120 hours) while maintaining 88.8% accuracy. This efficiency gain enables rapid processing of large datasets and frequent updates. The structured metadata output (Gap_Type, Specific_Missing_Training, Suggested_Intervention, Evidence_Quote) provides immediate analytical value beyond simple yes/no classification, while manual review typically produces less standardized or structured notes.

6. Discussion

6.1 Strengths

- TSB Structure Awareness: The 4-filter protocol's explicit handling of 'Causes' vs. 'Risk' distinction prevents false positives from incidental safety recommendations
- High Recall: 96.4% ensures comprehensive capture of training gaps, critical for research identifying preventable accidents
- Structured Extraction: Automatic generation of Gap_Type, Specific_Missing_Training, Suggested_Intervention, and Evidence_Quote enables systematic analysis

- Scalability: Two-stage pipeline processed 1,312 reports in ~10 minutes versus 80-120 hours for manual review

6.2 Limitations

- Moderate Precision: 75.0% precision indicates 25% of flagged cases require expert adjudication for borderline situations
- Explicit Evidence Requirement: The 'No Inference' filter may miss subtle implicit gaps requiring deeper contextual reasoning
- API Dependency: Rate limiting and quota management add implementation complexity
- Low Base Rate: Only 7.4% of original reports are training-related, indicating most GA accidents stem from other factors
- Dependence on TSB Documentation: The system can only identify training deficiencies explicitly documented within TSB investigation reports. Training issues that were present but not discussed by investigators cannot be identified by the classifier.

6.3 Research Implications

The 97 identified training-related accidents, enriched with structured metadata, enable systematic analysis of:

- Common patterns in training gaps across aircraft types, operation categories, and gap types (Missing Syllabus, SOP Omission, Fidelity Issues)
- Design of targeted training interventions using the Suggested_Intervention field to address the most prevalent systemic deficiencies

From an operational perspective, the identified accident reports provide an opportunity to examine recurring deficiencies in pilot knowledge, decision-making, procedural compliance, aircraft handling, and risk management. Rather than focusing only on individual accidents, the structured dataset enables researchers to identify systemic patterns that may inform improvements to flight training syllabus, recurrent proficiency training, and aviation safety education.

The future of Aviation 5.0 is expected to emphasize human-centric and sustainable innovation using technologies such as large language model (LLM)-based AI to address challenges related to aviation safety and sustainable workforce development (Bülbül et al., 2026). The current study demonstrates both the value and limitations of LLM-based automated accident classification, showing the need for future model refinement and continued human supervision.

7. Conclusions

The AI-powered two-stage classification system successfully processed 1,312 TSB aviation accident reports (1991-present), identifying 97 training-preventable incidents through rigorous filtering. Manual validation on 178 reports demonstrated strong performance: 88.8% accuracy, 75.0% precision, and 96.4% recall.

The system demonstrates significant advantages over traditional methods: 16× improvement in detection rate versus TSB website keyword search, 99% reduction in processing time versus manual review, and automatic generation of structured metadata enabling systematic analysis of training gaps.

While not intended to replace expert human judgment entirely, the system serves as a highly effective first-pass screening tool to reduce manual review workload while maintaining comprehensive capture of genuine training-related accidents. The high recall rate and structured output support the research objective of identifying systemic training deficiencies to inform targeted intervention design.

This approach provides a scalable framework in using artificial intelligence to support evidence-based improvements in pilot training, recurrent proficiency programs, and aviation safety research.

8. Recommendations

8.1 Model Refinement

- Implement active learning: Incorporate validated corrections to refine classification prompt and improve precision
- Develop confidence threshold policies: Auto-approve high-confidence cases (=10), flag borderline cases (7-9) for review
- Explore ensemble approaches: Combine multiple AI models for consensus-based classification

8.2 Research Application

- Conduct gap type analysis: Categorize the 97 cases by Gap_Type to identify most common deficiencies
- Design targeted interventions: Use Suggested_Intervention fields to develop specific training modules
- Cross-reference with regulations: Compare identified gaps against current CAR training requirements

8.3 Operational Deployment

- Establish periodic audit cycles: Monitor model drift and maintain classification quality
- Document edge cases: Build institutional knowledge of challenging classifications
- Expand to other jurisdictions: Apply methodology to NTSB, ATSB, and other safety databases

9. References

Government of Canada, Transportation Safety Board of Canada. (2024, July 19). *Air transportation safety investigations and reports - Transportation Safety Board of Canada*. https://www.tsb.gc.ca/eng/rapports-reports/aviation/index.html

Government of Canada, Transportation Safety Board of Canada. (2020, January 16). *About the TSB - Transportation Safety Board of Canada*. https://www.tsb.gc.ca/eng/qui-about/index.html

Nakanishi, M., & Ono, M. (2026). Datamining and modeling tacit knowledge of air accident investigators: Research aimed at improving efficiency and quality of investigations through ai. *Lecture Notes in Computer Science*, 116–126. https://doi.org/10.1007/978-3-032-12392-3_7

Nakanishi, M., Ono, M., & Tadaki, S. (2026). *Construction of an LLM-Based Analytical Model for Identifying Accident Factors and Their Causality from Factual Information in Aircraft Accident Investigation*. https://doi.org/10.2139/ssrn.6572622

Nanyonga, A., Turhan, U., & Wild, G. (2025). A Systematic Review of Machine Learning Analytic Methods for Aviation Accident Research. *Sci*, *7*(3), 124. https://doi.org/10.3390/sci7030124

Ziakkas, D., & Pechlivanis, K. (2023). Artificial intelligence applications in aviation accident classification: A preliminary exploratory study. *Decision Analytics Journal*, *9*, 100358. https://doi.org/10.1016/j.dajour.2023.100358

Dillon, R., Madsen, P., Holland, B., & Cao, D. (2024). How AI Can Help Learn Lessons from Incident Reporting Systems. In *2024 IEEE Aerospace Conference* (pp. 1–15). https://doi.org/10.1109/aero58975.2024.10521018

Bülbül, G., Niechwiej-Szwedo, E., Kearns, S., Cao, S., Barnett-Cowan, M., & Irving, E. (2026). Aviation 5.0: catalyst for workforce sustainability. *Aviation*, *30*(3), 198–216. https://doi.org/10.3846/aviation.2026.27635

10. Appendices

10.1 Air Taxi Inclusion Rationale

Air Taxi operations were included alongside pure General Aviation for the following reasons:

- Operational Similarity: Air Taxi operations use similar aircraft types as GA (Cessna 185, DHC-2 Beaver, Piper Navajo, Cessna 208 Caravan), operate from similar airports, and encounter comparable environmental challenges
- Training Pipeline Overlap: Air Taxi pilots receive their initial training through the same GA flight schools with identical Commercial Pilot License requirements
- Shared Training Challenges: Both groups face similar gaps in mountain flying, night operations in remote areas, single-pilot resource management, and emergency procedures in small aircraft

- Research Relevance: Air Taxi accidents often reveal systemic training deficiencies that are equally applicable to GA operations, such as insufficient type-specific emergency training or inadequate exposure to realistic operational conditions

Table 6. Composition of the Stage 1 filtered dataset by operation category.

Category	Count	% of Stage 1
Air Taxi / Air Taxi-related	277	46.9%
Pure General Aviation	314	53.1%
Total	591	100%

10.2 Classification Criteria and Case Examples

Training-Related (True Positive):

Accidents classified as training-related if TSB investigation explicitly identified preventable gaps:

- Missing Syllabus: 'Did not receive type-specific emergency procedures training'
- Manual/SOP Omission: 'Company had no procedure in place for...'
- Fidelity Issue: 'Night training conducted only in well-lit areas'

Non-Training-Related (True Negative):

- Pilot Error: Ignored warnings, poor judgment, intentional risk-taking
- Mechanical Failure: Engine failure, structural defect, maintenance issues
- Weather: Adverse conditions beyond pilot control
- External Factors: Bird strikes, ground vehicle collisions

10.3 Validation Accuracy by Confidence Score

The relationship between AI confidence scores and validation accuracy demonstrates that higher confidence correlates with higher classification accuracy:

Table 7. Validation accuracy by AI confidence score.

Confidence Score	Total Reports	Validated	Matches	Validation Accuracy
10	298	67	66	98.5%
9	282	100	84	84.0%
8	10	10	8	80.0%
≤7	5	1	0	0.0%

This pattern confirms that the AI's confidence metric provides useful signal for prioritizing human review: reports with confidence scores of 10 are highly reliable (98.5% accurate), while reports with scores ≤7 should receive mandatory expert verification.

10.4 Data Availability

Two Excel files accompany this report:

1. TSB_Full_Report_Links.csv

Contains all 1,312 original TSB aviation accident reports from 1991-present with columns: Year, Link (direct URL to TSB report)

2. TSB_Final_Training_Analysis.xlsx

Contains all 595 Stage 1 filtered reports (GA fixed-wing only) with complete AI classification results. Columns include: ReportID, is_systemic_gap, Gap_Type, Specific_Missing_Training, Suggested_Intervention, Evidence_Quote, Description, Confidence_Score, Aircraft_Type, Operation_Detail, Link, Year, and Manual Check validation results. Each report's Link field contains the direct URL to the full TSB investigation report on the TSB website (tsb.gc.ca).