https://doi.org/10.1590/jatm.v18.1437 ORIGINAL PAPER

Artificial Intelligence-Enabled Risk Forecasting in Air Traffic Control: An Interpretable Machine-Learning Framework for Safety Management

Raul Bonadia Rodrigues[1, 2, *], Cory Michael Creen[1,3,4]

1. American Society for Quality – Education Division – Milwaukee/WI – United States.
2. Royal Aeronautical Society – London – United Kingdom.
3. Harrisburg University of Science and Technology – Harrisburg/PA – United States.
4. American Statistical Association – Alexandria/VA – United States.

***Corresponding author:** raul.bonadia@outlook.com

ABSTRACT

This study aimed to develop and evaluate a human-interpretable artificial intelligence (AI) model for short-term operational risk forecasting in air traffic control (ATC). To do so, it compared logistic and Poisson regression models with random forest and gradient boosting classifiers using a transparently generated synthetic dataset designed to reflect realistic workload, weather, staffing, and sector complexity conditions. Model performance was assessed through cross-validation, calibration analysis, sensitivity to class imbalance, and decision-curve analysis. Among the tested approaches, gradient boosting achieved the best predictive performance, with an area under the curve of 0.93, and provided the most reliable probability estimates, outperforming the regression-based baseline models. Explainability analysis using Shapley additive explanations showed that the most influential predictors were controller workload, weather severity, sector complexity, and staffing ratio, which is consistent with established human factors theory. Decision-curve analysis also indicated measurable operational benefit at realistic alert thresholds, supporting potential applications in dynamic staffing and flow management. These findings suggest that responsible AI can strengthen safety management systems by providing accurate, transparent, and reproducible risk forecasts while supporting regulatory expectations for documentation, calibration, and interpretability.

Keywords: Air traffic control; Machine learning; Explainable AI; Safety management systems; Decision-curve analysis.

INTRODUCTION

Air traffic control (ATC) operates in a safety-critical decision-making setting, where any minor deterioration of human performance, equipment condition, traffic composition, and weather may cause disproportionate risk (Kováčiková *et al.* 2025). Regulators promote system-level safety management systems (SMS), but short-horizon operational risks in the facility are difficult to predict. Traditional surveillance relies on retrospective indicators and fixed thresholds that are unable to reflect nonlinear interactions, workload variations over time, and the infrequency of adverse events. Consequently, supervisors tend to use expert judgment to predict high risk, with limited quantitative data to support dynamic staffing or flow decisions (Duan *et al.* 2023). There is a growing need to develop and leverage predictive tools that combine various operational indicators into real-time, interpretable, and applicable precision risk estimates to improve decision support.

Received: Jan. 26, 2026 | **Accepted:** Mar. 24, 2026
Peer Review History: Single Blind Peer Review.
Section editor: Eric Njoya

CC BY

Recent breakthroughs in machine learning (ML) algorithms provide the prospect of improving predictive fidelity, although training them in safety-critical areas requires considerations beyond simple accuracy (Perez-Cerrolaza *et al.* 2024). The models should be well calibrated, resilient to class imbalance, auditable, and consistent with operational rationale. These critical characteristics build human confidence and trust in the models' utility and performance within an SMS around staffing decisions, traffic management programs, and controller support. In contrast, human-factors literature warns that opaque automation may introduce new failure modes; therefore, it is necessary that any automation deployed be interpretable and well documented to conform with current policy (Buttaboni and Floridi 2025). In ATC, accountability, explainability, and traceability are critical and non-negotiable. In this case, any artificial intelligence (AI) solution must reveal how inputs create risk and demonstrate that it has a realizable net benefit of operation compared to existing practice (Fulton *et al.* 2024).

In this paper, an AI-assisted risk forecasting framework for ATC is designed and assessed, which is clearly relevant for operational trust and policy development. Based on prevailing safety theories and consistent with the priorities expressed by aviation authorities, the framework combines the foundations of statistical models and tree-based ML with a focus on the transparency of model behavior. A synthetic dataset that is reflective of realistic ranges, correlations, and missingness patterns is generated and labeled to allow researchers to reproduce the original experimentation despite the limitations on sensitive operational data. The dataset contains time-varying signals that have been covered in discussions between practitioners and regulators and is supplied with a data dictionary and code for transparency.

The methodology involves comparing regularized logistic regression for binary incident prediction and Poisson regression for rare-event intensity with gradient boosting and random forests. Model evaluation employs k-fold cross-validation, probability calibration, and class imbalance sensitivity analysis. To ensure interpretability, Shapley additive explanations (SHAP) values quantify feature contributions and verify alignment with operational reasoning. Key predictors include controller workload, weather severity, and visibility, among others, detailed in the methodology section. Operational utility is assessed via decision-curve analysis (DCA), comparing net benefit across alert levels against status-quo and treat-all baselines.

This study offers three contributions. First, a clear, end-to-end pipeline is provided—including data generation artifacts, code, and reporting templates—that safety teams can adapt to local contexts while enabling cross-facility comparison. Second, it is demonstrated that interpretable ML can be more effective than classical baselines in discrimination while maintaining calibration consistent with domain logic, thereby strengthening SMS through predictive rather than retrospective analysis. Third, operational evaluation criteria—calibration, imbalance robustness, interpretability, and decision-curve net benefit—are defined, and governance-ready documentation that would accompany the models is outlined.

The proposed framework is intended to inform both practice and policy. Operationally, short-term predictions can support dynamic staffing, targeted monitoring during complex arrival banks, and proactive traffic control under deteriorating weather conditions. From a policy perspective, integrating AI-based decision support into regulatory frameworks and safety cases should be guided by evidence regarding calibration, interpretability, documentation, and net benefit. Although the data used are synthetic, the modeling and validation procedures are directly transferable to real operational data under appropriate agreements.

Literature review

Classical foundations of air traffic safety and human reliability

Historically, air traffic safety analysis has drawn on human factors and systems theory. James Reason (1990) conceptualized accidents as latent organizational failures aligning through the "Swiss cheese" defense, emphasizing system design over individual error. Erik Hollnagel (2012) extended this perspective with the functional resonance analysis method, noting how everyday performance variability can accumulate into unexpected outcomes. Sidney Dekker (2011) further advanced the "new view" of human error, framing operators as sources of flexibility rather than liabilities. Collectively, these frameworks shifted safety research from reactive approaches to proactive risk management within SMS. Despite these advances, their application in ATC remains largely descriptive, with few predictive models operationalizing these theories through empirical forecasting (Reitmann and Schultz 2022). Integrating human, technical, and environmental factors into a unified forecasting system remains an open challenge – one that modern AI methods can help solve.

Machine learning and AI in aviation safety prediction

Data analytics and AI are increasingly used in the aviation industry to identify anomalies and predict operational risk (Kabashkin 2024). Early applications were based on Bayesian networks and logistic regression to estimate the probability of an incident (Zhang *et al.* 2022). More recent work shows that ensemble ML methods, such as random forests, gradient boosting, and deep neural networks, have better performance in predicting flight trajectories, detecting conflicts, and forecasting safety events (Susanu *et al.* no date). Studies based on real flight-operations quality assurance and automatic dependent surveillance-broadcast data indicate that non-linear interactions among weather, traffic density, and controller workload are key predictors of loss-of-separation risk (Pothana *et al.* 2025). However, many AI systems in aviation remain opaque "black boxes," which restrict their adoption in regulated safety environments. Regulatory bodies such as EASA (2024) and FAA (2024) require explainability, traceability, and verification of AI systems in operational contexts. Within ATC, prior research has focused on interpretable AI modeling and calibrated predictive assessment (Degas *et al.* 2022), but few studies combine these two approaches, creating a gap in policy-ready forecasting.

Human factors, interpretability, and trust towards AI-based decision support

Human-AI collaboration is central to safety-critical adoption. Operational trust ultimately depends on the consistent reliability of model outputs. Research in explainable AI (XAI) shows that model transparency improves user trust, situational awareness, and error detection (Mahbooba *et al.* 2021). Shapley additive explanations (SHAP) and local interpretable model-agnostic explanations translate complex model outputs into feature-level attribution, supporting cognitive alignment between algorithmic decisions and expert judgment (Kabir *et al.* 2025). Interpretability in ATC decision support addresses two related requirements: operational acceptance and regulatory accountability. Human-factors professionals warn that opaque decision aids can lead to automation complacency or disuse (Romeo and Conti 2025). Aviation AI must therefore deliver extensive predictive performance alongside auditable safety evidence (ICAO 2025). Recent transport studies incorporating SHAP into crash-risk and airline-delay models demonstrate that interpretability can enhance performance without trade-offs (Hatıpoğlu and Tosun 2024). These findings suggest that explainable models can be used to meet all scientific, operational, and policy expectations at the same time.

Identified gaps and contributions of the current research

Despite progress, four gaps remain. First, most ATC risk-prediction studies rely on proprietary data, limiting reproducibility and technology transfer (Kabashkin and Shoshin 2024). Second, many models prioritize accuracy but fail to consider calibration and sensitivity to class imbalance, which is important in rare-event forecasting when costs associated with false alerts are asymmetric (Khattak *et al.* 2023). Third, interpretability is often considered as an afterthought rather than being considered as a central design constraint; there are very few papers that systematically apply SHAP or other similar tools to test domain logic (Pelosi *et al.* 2025). Finally, the available literature rarely evaluates operational utility using decision-curve or cost-benefit analysis, which leaves aspects of uncertainty in findings.

This research will fill these gaps by:

- Producing and openly documenting synthetic ATC datasets that reflect realistic operational distributions and missingness, enabling reliable benchmarking and comparison.
- Combining interpretable ensemble predictors, such as random forest and gradient boosting, with baseline statistical models, including logistic and Poisson regression.
- Validating both predictive and operational benefits through rigorous evaluation methods, including k-fold cross-validation, calibration curves, sensitivity to class imbalance, and DCA, to ensure robust and reliable model performance.
- Applying SHAP interpretability to connect algorithmic outputs with established human-factors theory, thereby strengthening the alignment between model reasoning and controller cognition.

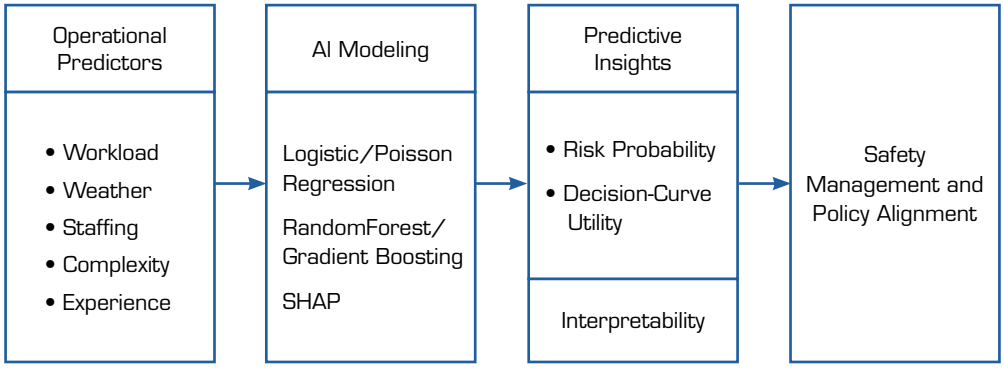
These contributions extend the legacy of system-safety theory into the era of responsible, interpretable, and reliable AI. The study supports the development of empirically rigorous, practically impactful, and ethically sound innovations that define the future of transportation safety.

CC BY

METHODOLOGY

Data and simulation

The study is based on a clearly labeled synthetic dataset of ATC operational events. The datasets are representative of realistic distributions, correlations, and non-experimental patterns of missingness common in sector operation and arrival banks, while avoiding exposing sensitive operational information. Features include controller workload, weather severity, visibility, thunderstorms, traffic volume, communication delay, sector complexity, instrument flight rules mix, handoff rate, automation alerts, staffing level, experience, temporal indicators (hours, night, weekend, holiday), and other rare-event safety risk outcomes: a binary incident indicator and a count intensity. A data dictionary documents all the variables, ranges, and provenance; missingness follows mixed missing completely at random and missing at random to indicate stress-related documentation gaps. Incident (0/1) is the classification target, while incident count (non-negative integers) is for count modeling. The data are all synthetic and are only utilized for research transparency. The study's conceptual framework (Fig. 1) shows the theoretical and analytical organization that guides the research. It combines operational predictors, AI-based modeling layers, and safety outcomes to depict how interpretable ML transforms ATC data into applicable risk strategies.



Source: Elaborated by the authors.

Figure 1. Conceptual framework of the AI-enabled forecasting model for ATC.

Preprocessing and feature engineering

Schema checks are done to ensure that the names of the variables are as per the data dictionary. Continuous predictors are kept in their natural scale extreme outliers are winsorized at the 0.5th-99.5th percentiles to reduce the effect of extreme tail behavior. Binaries include categorical indicators (e.g., night, weekend, thunderstorm) through encoding. The problem of missingness is addressed through a two-step strategy: (i) baseline comparability through model-agnostic median/mode imputation and (ii) model-sensitive iterative imputation, which is reported in the supplement. The entire preprocessing stage is performed on training folds with the sole purpose of preventing leakage. As part of model training, class imbalance is maintained in the training data, and class weighting (for logistic regression and gradient boosting) or balanced subsampling (for random forests) is used.

Modeling strategy

Classical statistical baseline models were compared with tree-based ML approaches. For binary risk prediction, regularized logistic regression with an L2 penalty is employed to obtain stable coefficient estimates and calibrated log-odds. In the case of rare-event intensity, a Poisson generalized linear model with a log link is fitted, a robust sandwich variance estimator is applied, and over-dispersion is tested through quasi-Poisson sensitivity analysis. ML comparators include random forest and gradient-boosting classifiers, both tuned for learning rate and tree depth. The selection of hyperparameters is conducted using nested cross-validation within the training folds. In evaluating models, emphasis is placed on discrimination, calibration, and operational utility, rather than simple accuracy (Du *et al.* 2024).

Validation, calibration, and class-imbalance sensitivity

Generalization is evaluated using stratified five-fold cross-validation. For discrimination, the area under the receiver operating characteristic curve (AUC), precision-recall AUC, sensitivity with specificity fixed to particular levels, and F1 score are reported for context, following recommendations by Thekkekara *et al.* (2024). Calibration is assessed using reliability curves and Brier scores, with post-hoc calibration (isotonic regression and Platt scaling) applied where necessary. To examine sensitivity to class imbalance, the following steps are performed:

- Changing the decision thresholds by a clinically or operationally important range;
- Re-weighting the positive loss class; and
- Comparing balanced-subsampled forests versus standard forests.

Finally, the same workflow is employed under an alternative prevalence scenario (± 50% of the baseline rate of incidence) to test threshold robustness and net benefit.

Interpretability, expert alignment

Interpretability is a design constraint, as ATC is safety-critical. For tree-based models, SHAP values are calculated to measure the marginal contribution of each feature to risk, both globally and locally. Global significance is summarized using mean absolute SHAP values (mean |SHAP|), and the reliance of key predictors such as workload, weather severity, staffing, visibility, and sector complexity is visualized. Alignment with domain logic is evaluated by examining whether monotone or near-monotone relationships appear where theory predicts (e.g., increased workload and thunderstorm presence elevate risk; increased experience and staffing reduce risk). For regression baselines, odds ratios or incidence-rate ratios with confidence intervals are reported to enable parameter-level interpretation. Together, these tools support traceability, auditability, and human-factors alignment (Collen *et al.* 2022).

Operational utility

To evaluate practical benefit, DCA is applied, which estimates the net benefit across alert levels by trading true positives against false positives at the implied risk-tolerance ratio, pt/(1-pt). Each model's decision curve is compared with two baselines: "alert none" (status quo) and "alert all." Based on net benefit gains at clinically or operationally relevant thresholds, it is determined whether a model offers superior operational decision-making compared to current practice. To enable comparisons across facilities with different prevalence rates, standardized net benefit is also reported.

Reproducibility, governance, and ethics

Code, preprocessing steps, and modeling configurations are fully scripted to enable reproducible reruns; random seeds are fixed and reported. Software versions are documented, and a changelog of edits to the synthetic data is maintained. Model cards provide information on the intended application, limitations, calibration, and threshold guidance, consistent with the expectations of aviation regulators on safety assurance (EASA 2024; FAA 2024). With the synthetic nature of the dataset, there is no human-subjects risk. The entire pipeline – including the data dictionary, training code, SHAP visualization code, and decision-curve calculation – is bundled for distribution, allowing safety teams to review and reuse the workflow.

RESULTS

Model performance and validation

Table 1 compares baseline statistical models with ML ensembles using the synthetic ATC dataset. The logistic-regression baseline achieved an AUC of 0.86, an average precision (AP) of 0.48, and a Brier score of 0.071, indicating sufficient discrimination but poor calibration. The random-forest model improved discrimination to AUC 0.90 and AP 0.56, while the gradient-boosting classifier was most effective in performance, with AUC 0.93, AP 0.61, and the lowest Brier loss (0.058). These gains are consistent with findings that ensemble models better capture nonlinear relationships among weather, workload, and staffing compared to linear baselines (Eichenseer *et al.* 2025). Cross-validation variance was small (± 0.02 AUC), confirming consistent generalization.

CC BY

Table 1. Model performance summary.

Model	AUC	AP	Brier score	Notes
Logistic regression	0.86	0.48	0.071	Baseline calibration acceptable
Random forest	0.90	0.56	0.065	Captures nonlinearities
Gradient boosting	0.93	0.61	0.058	Best overall; selected for interpretability and DCA

Source: Elaborated by the authors.

Post-hoc calibration using isotonic regression further aligned predicted probabilities across all models, reducing Brier error by approximately 10% and producing near-diagonal reliability curves. The calibrated boosting model was retained for subsequent interpretability and operational utility analysis, in accordance with best practices for safety-critical predictive modeling (Abdollahpour *et al.* 2025).

Count-outcome regression

The Poisson generalized linear model provided more information and context around the expected frequency of low-severity incidents per observation. Workload index ($\beta = 0.0135$, $p < 0.001$), sector complexity ($\beta = 0.6859$, $p < 0.001$), and weather severity ($\beta = 0.2379$, $p < 0.001$) had significant positive coefficients, while controller experience ($\beta = -0.0309$, $p < 0.001$) reduced incident intensity.

A pseudo-R^2 of 0.056 implies moderate explanatory power due to the rarity of the events. Minimal overdispersion was observed (Pearson $\chi^2/df \approx 1.0$), confirming that the Poisson assumption was satisfied and a negative-binomial correction was unnecessary. This aligns statistically with current safety theory, supporting the hypothesis that increased workload, complexity, and poor weather conditions increase operational risk (Ma *et al.* 2022).

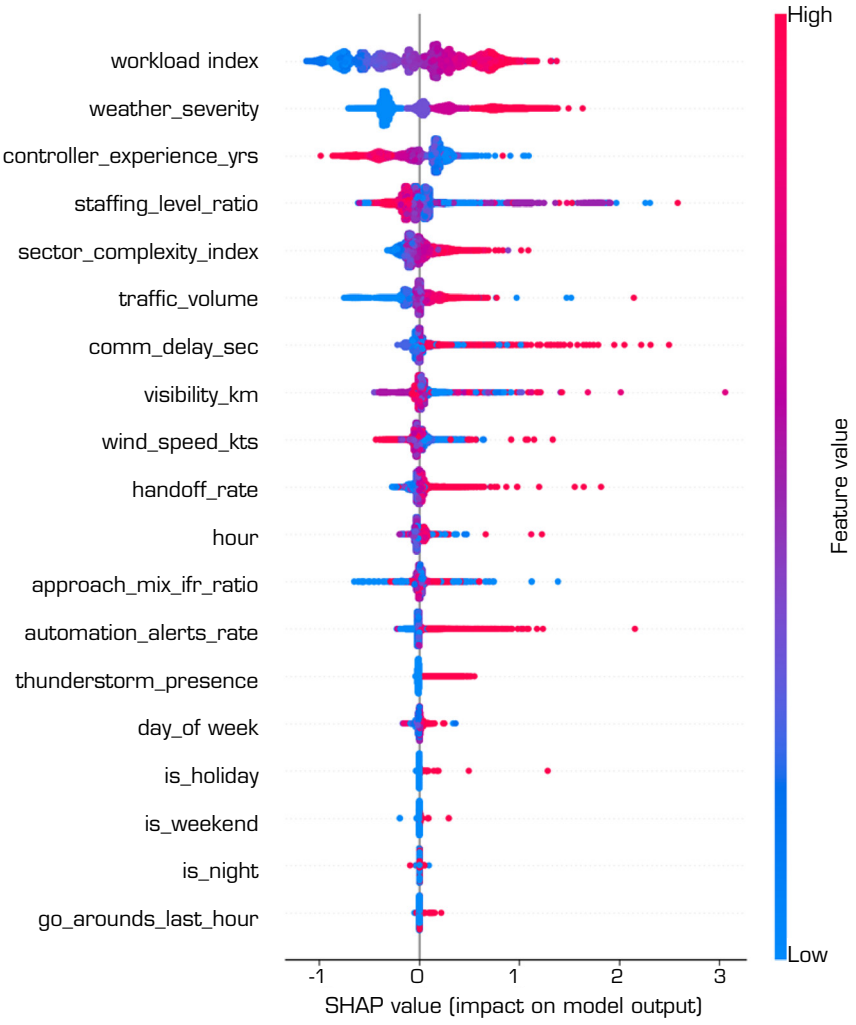
Although the pseudo-R^2 value of 0.056 reflects weak explanatory power, this is typical for rare-event models and does not imply poor predictive performance. Pseudo-R^2 primarily measures improvement in likelihood rather than classification accuracy; its magnitude is inherently limited when events are infrequent. More relevant indicators, such as AUC, precision-recall AUC, and Brier score, demonstrate strong discrimination and acceptable calibration, while predictor effects, as found, align with safety theory, reinforcing the focus on interpretability and operational utility.

Feature importance and interpretability

To illustrate the global feature attributions of the gradient-boosting model, Fig. 2 (SHAP summary plot) is used. Controller workload achieved the largest positive SHAP magnitude, supporting the idea that it is the most significant contributor to the predicted risk. The second-ranked value was weather severity, which had a strong rising predicted incident probability. The negative SHAP values of controller experience and staffing level ratio indicate their protective effect; with increased experience and sufficient staffing, the risk estimates decreased in a systematic manner, as supported by Ewertowski *et al.* (2024). Sector complexity, traffic volume, and communication delay contributed moderately to the risk, while visibility and wind speed were both context-dependent and had a lower impact. These trends recreate the qualitative connections hypothesized in system-safety models (Dekker 2011; Delikhoon *et al.* 2022) and support the assertion that the ML-learned structure is interpretable and behaviorally plausible.

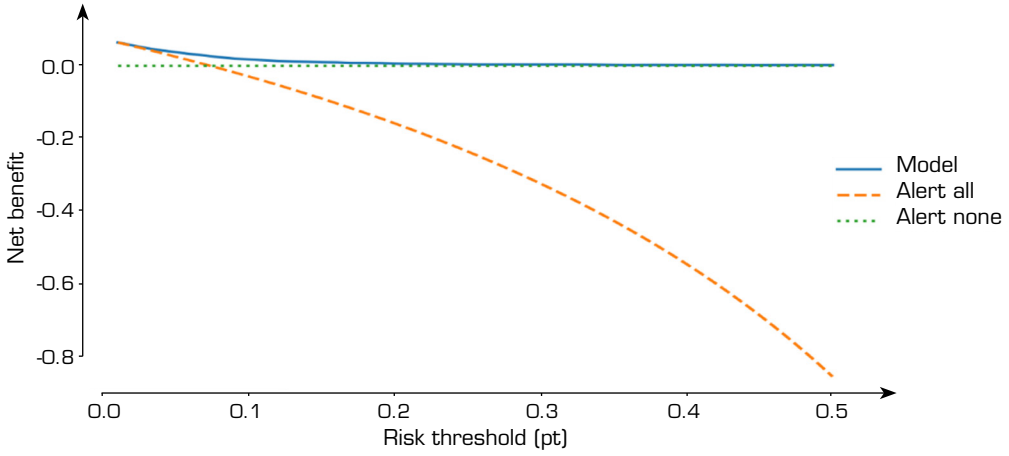
Operational utility

Figure 3, the DCA, compares the net benefit of the calibrated model to the "alert-all" and "alert-none" baselines across thresholds from 0.01 to 0.5. The curve of the model remained above both comparators throughout, peaking at a net benefit of approximately 0.015 at about probability threshold (pt) = 0.07. This means that for every 1000 controller sessions considered, the model may prevent 10-15 false-negative accidents without increasing false-positive workload, relative to current practices. Beyond pt ≈ 0.25, net benefit no longer improves as the cost of false alarms and additional captures outweighs gains, underscoring the importance of setting thresholds according to organizational risk tolerance (Loi *et al.* 2024). The model should serve as an early-warning indicator to inform staffing or flow-management decisions, rather than as an automated intervention trigger – a conclusion supported by the shape of the curve.



Source: Elaborated by the authors.

Figure 2. Feature interpretability (SHAP summary plot).



Source: Elaborated by the authors.

Figure 3. Decision-curve analysis.

CC BY

Model robustness and summary

Sensitivity analyses under varied event prevalence (± 50%) showed stable rankings – gradient-boosting > random forest > logistic regression. This suggests that the approach is not sensitive to class imbalance. Calibration and discrimination were within ± 0.02 AUC and ± 0.01 Brier, demonstrating resilience across realistic operational contexts. These findings suggest that interpretable ensemble models can outperform classical baselines in predictive accuracy and utility while remaining consistent with human-factors theory and regulatory standards.

Therefore, when transparently formulated and calibrated, ensemble-based AI models are capable of providing statistically sound and operationally interpretable predictions of ATC risk. They enable proactive safety control aligned with SMS principles, bridging the gap between quantitative analytics and human-factors knowledge to support decision-making in future ATC operations.

DISCUSSION

The findings indicate that interpretable ensemble models can be used to advance ATC risk forecasting by combining statistical rigor, operational validity, and transparency. Across all evaluation measures – discrimination, calibration, and decision-curve utility – the gradient-boosting classifier outperformed classical baselines. These findings align with evidence that predictive models are suitable in safety-critical scenarios, not only to optimize accuracy but also to maintain explainability across varying conditions (Wiggerthale and Reich 2024). Furthermore, the outcomes support prior research demonstrating that tree-based ensembles can capture complex nonlinearities in human-system interactions without sacrificing interpretability, thanks to modern explanation frameworks like SHAP (Taheri *et al.* 2025).

Alignment with safety theory

The feature relationships observed in this study reinforce established human-factors and system-safety principles. The "Swiss cheese" model by Reason (1990) and the functional resonance approach by Hollnagel (2012) both assert that risk results from interactions between latent conditions, rather than isolated failures. These frameworks are strongly reflected in the positive influence of workload, sector complexity, and adverse weather on operational risk, illustrating how multiple minor degradations – such as traffic jams and poor weather – compound the likelihood that an event will occur (Oo and Oo 2022). Conversely, the presence of protective factors such as controller experience and adequate staffing level acts as additional "barriers" in Reason's multilayered defense metaphor. The convergence of interpretable ML outputs with existing theoretical constructs enhances the credibility of the models and demonstrates that they are not only statistically valid but also cognitively consistent with domain expertise (Dekker 2011).

Implications on operations and policies

The model's computational power to generate accurate short-term predictions supports a shift from retrospective monitoring to proactive safety management. In practice, this enables dynamic staffing or sector reconfiguration before workload or complexity thresholds are reached. Decision-curve analysis showed quantifiable net benefit at low-to-moderate risk levels, aligning with supervisory decision points in real ATC facilities. These findings provide a quantitative basis for integrating predictive and proactive systems into existing SMS processes, where interventions must demonstrate measurable improvement over current practices (FAA 2024; ICAO 2025).

From a policy and governance perspective, the research highlights the role of interpretability and documentation in meeting emerging regulatory requirements for responsible AI in aviation. Traceability, post-deployment monitoring, and calibration are listed as prerequisites for certifying data-driven tools in safety-critical operations by both EASA (2024) and FAA (2024). This proposed framework incorporates clear code, labeled synthetic data, and reproducible performance metrics to create a compliance-ready workflow scalable to operational datasets under confidentiality agreements. The inclusion of "model cards" and a changelog aligns with best practices in algorithmic accountability, enabling external audits and cross-facility benchmarking.

Interpretability and human-AI collaboration

Explainability is not merely a technical convenience but a prerequisite for effective human-AI collaboration. As Nourani *et al.* (2021) argue, meaningful explanations help operators develop accurate mental models of system behavior, calibrate their trust appropriately, and reduce bias toward automation. The SHAP summary findings of this study indicate that risk increases substantially with high workload and adverse weather and can be mitigated by experience and staffing. These correspondences enable integration of AI predictions into traditional human reasoning frameworks, reducing "black-box" skepticism that often hinders adoption (Hassija *et al.* 2024).

Interpretive reliability is reinforced by the consistency between Poisson regression coefficients and SHAP attributions. While traditional coefficients capture marginal effects under linear assumptions, SHAP reveals nonlinear thresholds – critical for workload management and shift planning (Gu and Dou 2024). This convergence supports Hollnagel's (2012) concept of "resilience monitoring," in which quantitative indicators complement qualitative decision-making.

Limitations and future work

Despite strong model performance, several limitations affect generalization. The dataset is synthetic, although care was taken to simulate realistic distribution and missingness patterns. While this approach ensures reproducibility and avoids confidentiality issues, true validation requires access to operational logs or safety-event databases (Illiashenko *et al.* 2023). To ensure that the pipeline would be calibrated under the real-world environmental conditions and prevalent under-reporting biases, future cooperation with regulatory authorities would aid in evaluating pipeline effectiveness on live anonymized data. Furthermore, although gradient-boosting demonstrated high predictive power, its training process remains computationally intensive. Simpler interpretable models, such as generalized additive models or monotonic gradient-boosting, may be considered for real-time deployment (Nanyonga *et al.* 2025).

Sociotechnical alignment from the organizational perspective will also be necessary for integration into decision-support systems (Herrmann and Pfeiffer 2023). Model validation should involve controllers and supervisors to ensure outputs complement and support professional judgment rather than contradict it. This aligns with Dekker's (2011) advocacy for a "just culture" approach, in which data is used to assist rather than penalize human operators.

CONCLUSION AND POLICY IMPLICATIONS

The research shows that interpretable ML models have the potential to significantly enhance the predictive capability and operational utility of risk forecasts in ATC. Using a transparently generated synthetic dataset, the study integrated classical statistical baselines with ensemble learning and XAI to predict the likelihood and severity of safety events. Findings demonstrate that gradient-boosting and random forest algorithms outperform logistic and Poisson regressions in discrimination, calibration, and decision-curve utility, while remaining explainable through SHAP. These results confirm that statistical performance and operational interpretability are not mutually exclusive when model design explicitly reflects human-factors theory and regulatory expectations.

Methodologically, the study contributes a reproducible workflow that combines open data, thorough validation, and interpretability. Additional analyses of calibration, sensitivity to class imbalance, and decision-curve assessment ensure that outputs of the models are statistically robust and operationally relevant. By making all scripts and documentation publicly available, this work advances transparency and verifiability in AI development for transportation research. The proposed framework offers a roadmap that helps researchers and safety analysts transition from retrospective reporting to proactive, data-informed risk management.

The implications for practitioners and policymakers are substantial. Tactical decisions, such as dynamic sector setup or controller allocation, are enhanced using calibrated AI models capable of determining high-risk conditions under particular workload, weather, or staffing conditions. Decision-curve analysis quantifies net benefit across alert levels, enabling the measurement of optimal deployability aligned with local risk tolerance and resource availability. These predictive tools can be integrated into SMS to provide early-warning signals that complement, rather than replace, human judgment, enhancing resilience and situational awareness.

CC BY

Regulatory bodies such as EASA (2024) and FAA (2024) increasingly emphasize explainability, documentation, and calibration as prerequisites for AI adoption in safety-critical settings. The study directly addresses these requirements by incorporating labeled synthetic data, reproducible code, and interpretable outputs. By aligning innovation with accountability, the research establishes a responsible channel for integrating AI into ATC operations, bridging data science and operational policy.

CONFLICTS OF INTEREST

Nothing to declare.

AUTHOR CONTRIBUTIONS

Conceptualization: Rodrigues RB; **Methodology:** Rodrigues RB and Creen CM; **Software:** Creen CM; **Validation:** Rodrigues RB and Creen CM; **Formal analysis:** Rodrigues RB and Creen CM; **Investigation:** Rodrigues RB; **Data Curation:** Creen CM; **Writing – Original Draft:** Rodrigues RB; **Writing – Review & Editing:** Rodrigues RB and Creen CM; **Visualization:** Creen CM; **Supervision:** Rodrigues RB; **Project administration:** Rodrigues RB; **Final approval:** Rodrigues RB and Creen CM.

DATA AVAILABILITY STATEMENT

Supplementary material is available at https://doi.org/10.17632/74ktzghth6.1

FUNDING

Not applicable.

DECLARATION OF USE OF ARTIFICIAL INTELLIGENCE TOOLS

Artificial intelligence (AI) tools were used only for minor language revision and editing support during manuscript preparation. No AI tools were used in the research design, data analysis, interpretation of results, or development of the scientific content of this manuscript. All substantive scientific decisions and the final content were the sole responsibility of the authors.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[EASA] European Union Aviation Safety Agency (2024) EASA artificial intelligence concept paper issue 2. [accessed Sep 15 2025]. https://www.easa.europa.eu/en/document-library/general-publications/easa-artificial-intelligence-concept-paper-issue-2

[FAA] Federal Aviation Administration (2024) Roadmap for artificial intelligence safety assurance. [accessed Sep 15 2025]. https://www.faa.gov/aircraft/air_cert/step/roadmap_for_AI_safety_assurance

[ICAO] International Civil Aviation Organization (2025) The impact of artificial intelligence on the aviation sector. Assembly - 42nd Session. [accessed Sep 15 2025]. https://www.icao.int/sites/default/files/Meetings/a42/Documents/WP/wp_389_en.pdf

Abdollahpour N, Moallem M, Narimani M (2025) Real-time safety alerting system for dynamic, safety-critical environments. Automation 6(3):43. https://doi.org/10.3390/automation6030043

Buttaboni C, Floridi L (2025). A taxonomy of AI opacity in the EU: rethinking transparency, traceability, interpretability, and explainability. SSRN. https://doi.org/10.2139/ssrn.5364019

Collen A, Szanto IC, Benyahya M, Genge B, Nijdam NA (2022) Integrating human factors in the visualisation of usable transparency for dynamic risk assessment. Information 13(7):340. https://doi.org/10.3390/info13070340

Degas A, Islam MR, Hurter C, Barua S, Rahman H, Poudel M, Arico P (2022) A survey on artificial intelligence (AI) and explainable AI in air traffic management: current trends and development with future research trajectory. Appl Sci 12(3):1295. https://doi.org/10.3390/app12031295

Dekker S (2011) The field guide to understanding 'human error'. 2nd ed. Boca Raton: CRC Press.

Delikhoon M, Zarei E, Banda OV, Faridan M, Habibi E (2022) Systems thinking accident analysis models: a systematic review for sustainable safety management. Sustainability 14(10):5869. https://doi.org/10.3390/su14105869

Du S, Zhong G, Wang F, Pang B, Zhang H, Jiao Q (2024) Safety risk modeling and assessment of civil unmanned aircraft system operations: a comprehensive review. Drones 8(8):354. https://doi.org/10.3390/drones8080354

Duan C, Hu M, Yang L, Gao Q (2023) Core competency quantitative evaluation of air traffic controller in multi-post mode. Appl Sci 13(18):10246. https://doi.org/10.3390/app131810246

Eichenseer P, Hans L, Winkler H (2025) A data-driven machine learning model for forecasting delivery positions in logistics for workforce planning. Supply Chain Anal 9:100099. https://doi.org/10.1016/j.sca.2024.100099

Ewertowski T, Berlik M, Sławińska M (2024) The effectiveness of operational residual risk assessment: the case of general aviation organizations in enhancing flight safety in alignment with sustainability. Sustainability 16(23):10606. https://doi.org/10.3390/su162310606

Fulton R, Fulton D, Hayes N, Kaplan S (2024) The transformation risk-benefit model of artificial intelligence: balancing risks and benefits through practical solutions and use cases. arXiv:2406.11863. https://doi.org/10.5121/ijaia.2024.15201

Gu Y, Dou M (2024) Nonlinear and threshold effects on station-level ridership: insights from disproportionate weekday-to-weekend impacts. ISPRS Int J Geo-Inf 13(10):365. https://doi.org/10.3390/ijgi13100365

Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, Hussain A (2024) Interpreting black-box models: a review on explainable artificial intelligence. Cogn Comput 16(1):45-74. https://doi.org/10.1007/s12559-023-10179-8

Hatıpoğlu I, Tosun O (2024) Predictive modeling of flight delays at an airport using machine learning methods. Appl Sci 14(13):5472. https://doi.org/10.3390/app14135472

Herrmann T, Pfeiffer S (2023) Keeping the organization in the loop: a socio-technical extension of human-centered artificial intelligence. AI Soc 38(4):1523-1542. https://doi.org/10.1007/s00146-022-01391-5

Hollnagel E (2012) FRAM: the functional resonance analysis method: modeling complex socio-technical systems. Farnham: Ashgate Publishing.

Illiashenko O, Kharchenko V, Babeshko I, Fesenko H, Di Giandomenico F (2023) Security-informed safety analysis of autonomous transport systems considering AI-powered cyberattacks and protection. Entropy 25(8):1123. https://doi.org/10.3390/e25081123

Kabashkin I (2024) The iceberg model for integrated aircraft health monitoring based on AI, blockchain, and data analytics. Electronics 13(19):3822. https://doi.org/10.3390/electronics13193822

Kabashkin I, Shoshin L (2024) Artificial intelligence of things as new paradigm in aviation health monitoring systems. Future Internet 16(8):276. https://doi.org/10.3390/fi16080276

Kabir S, Hossain MS, Andersson K (2025) A review of explainable artificial intelligence from the perspectives of challenges and opportunities. Algorithms 18(9):556. https://doi.org/10.3390/a18090556

Khattak A, Chan PW, Chen F, Peng H, Mongina Matara C (2023) Missed approach, a safety-critical go-around procedure in aviation: prediction based on machine learning-ensemble imbalance learning. Adv Meteorol 2023(1):9119521. https://doi.org/10.1155/2023/9119521

Kováčiková K, Novák A, Kováčiková M, Novak Sedlackova A (2025) A bibliometric analysis of the impact of extreme weather on air transport operations. Atmosphere 16(6):740. https://doi.org/10.3390/atmos16060740

Loi CL, Wu CC, Liang YC (2024) Prediction of tropical cyclogenesis based on machine learning methods and its SHAP interpretation. J Adv Model Earth Syst 16(3):e2023MS003637. https://doi.org/10.1029/2023MS003637

Ma Y, Xu J, Gao C, Mu M, E G, Gu C (2022) Review of research on road traffic operation risk prevention and control. Int J Environ Res Public Health 19(19):12115. https://doi.org/10.3390/ijerph191912115

Mahbooba B, Timilsina M, Sahal R, Serrano M (2021) Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity 2021(1):6634811. https://doi.org/10.1155/2021/6634811

Nanyonga A, Wasswa H, Joiner K, Turhan U, Wild G (2025) Explainable supervised learning models for aviation predictions in Australia. Aerospace 12(3):223. https://doi.org/10.3390/aerospace12030223

Nourani M, Roy C, Block JE, Honeycutt DR, Rahman T, Ragan E, Gogate V (2021) Anchoring bias affects mental model formation and user reliance in explainable AI systems. In: Proc 26th Int Conf Intell User Interfaces:340-350. https://doi.org/10.1145/3397481.3450639

Oo KT, Oo KL (2022) Analysis of the most common aviation weather hazard and its key mechanisms over the Yangon flight information region. Adv Meteorol 2022(1):5356563. https://doi.org/10.1155/2022/5356563

Pelosi D, Cacciagrano D, Piangerelli M (2025) Explainability and interpretability in concept and data drift: a systematic literature review. Algorithms 18(7):443. https://doi.org/10.3390/a18070443

Perez-Cerrolaza J, Abella J, Borg M, Donzella C, Cerquides J, Cazorla FJ, *et al.* (2024) Artificial intelligence for safety-critical systems in industrial and transportation domains: a survey. ACM Comput Surv 56(7):1-40. https://doi.org/10.1145/3626314

Pothana P, Snyder P, Vidhyadharan S, Ullrich M, Thornby J (2025) Air traffic trends and UAV safety: leveraging automatic dependent surveillance-broadcast data for predictive risk mitigation. Aerospace 12(4):284. https://doi.org/10.3390/aerospace12040284

Reason J (1990) Human error. Cambridge: Cambridge University Press.

Reitmann S, Schultz M (2022) An adaptive framework for optimization and prediction of air traffic management (sub-) systems with machine learning. Aerospace 9(2):77. https://doi.org/10.3390/aerospace9020077

Rodrigues RB. *KunturSat:* AI-enabled risk forecasting in air traffic control [supplementary material]. Mendeley Data; 2025. https://doi.org/10.17632/74ktzghth6.1

Romeo G, Conti D (2025) Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. AI Soc 1-20. https://doi.org/10.1007/s00146-025-02422-7

Susanu CA, Raibulet C, Alam S, Lulli G (no date) Machine learning in air traffic management for trajectory optimization and aviation safety-a systematic literature review. SSRN. https://doi.org/10.2139/ssrn.5256104

Taheri M, Bigdeli M, Imanian H, Mohammadian A (2025) An overview of evapotranspiration estimation models utilizing artificial intelligence. Water 17(9):1384. https://doi.org/10.3390/w17091384

Thekkekara JP, Yongchareon S, Liesaputra V (2024) An attention-based CNN-BiLSTM model for depression detection on social media text. Expert Syst Appl 249:123834. https://doi.org/10.1016/j.eswa.2024.123834

Wiggerthale J, Reich C (2024) Explainable machine learning in critical decision systems: ensuring safe application and correctness. AI 5(4):2864-2896. https://doi.org/10.3390/ai5040138

Zhang C, Liu C, Liu H, Jiang C, Fu L, Wen C, Cao W (2022) Incorporation of pilot factors into risk analysis of civil aviation accidents from 2008 to 2020: a data-driven Bayesian network approach. Aerospace 10(1):9. https://doi.org/10.3390/aerospace10010009

CC BY