World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

Predictive training analytics for competency-based pilot instruction: Early Detection of Checkride Risk, Procedural Noncompliance and Safety-Relevant Learning Gaps

Okebu Daniel Tobechukwu Chukwuemeka [1] and Munashe Naphtali Mupa [2,*]

[1] *Embry Riddle Aeronautical University.*
[2] *Hult International Business School.*
Okebu Daniel Tobechukwu Chukwuemeka; ORCiD: 0009-0001-5409-2903
Munashe Naphtali Mupa; ORCiD: 0000-0003-3509-861X

World Journal of Advanced Research and Reviews, 2026, 30(03), 1928-1942

Publication history: Received on 15 May 2026; revised on 24 June 2026; accepted on 26 June 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.3.1774

Abstract

This paper develops a predictive training analytics framework for competency-based pilot instruction. The objective is to detect checkride risk, procedural noncompliance, and safety-relevant learning gaps early enough for instructors, chief instructors, and training managers to intervene before deficiencies become repeated stage-check failures, unsafe normalization of deviations, or weak practical-test readiness. The study combines public aviation safety sources, including the National Transportation Safety Board aviation accident database, the NASA Aviation Safety Reporting System, FAA Airman Certification Standards, FAA risk-management and aviation-instructor guidance, and Part 61/Part 141 regulatory structures, with a reproducible synthetic training-record dataset representing variables commonly available to flight schools. Because individual student pilot lesson records and practical-test outcomes are not publicly released at scale, the synthetic dataset is used only as a transparent modeling demonstration that can be replaced with real school records in implementation.

The analytical design compares logistic regression, random forest, and gradient-boosting classifiers using lesson-stage variables such as repeat lessons, unsatisfactory maneuvers, procedural-deviation rate, radio-readback errors, IFR task incompletion, approach-stability violations, weather and workload exposure, instructor interventions, safety-event flags, knowledge-test score, stage-check score, lesson gaps, instructor-comment risk terms, and checklist-adherence scores. The logistic-regression model produced the strongest test performance in this simulation, with a test AUC of 0.837 and a five-fold cross-validation AUC of 0.850. The framework is not proposed as a punitive student-ranking system; it is proposed as a non-punitive, safety-management-aligned decision support tool that converts training records into early warning indicators, debrief priorities, proficiency gates, and individualized remediation.

Keywords: Pilot Training Analytics; Checkride Risk; Competency-Based Training; FAA Airman Certification Standards; Procedural Noncompliance; Stage Checks; Machine Learning; Aviation Safety; Part 141; Part 61; Learner-Centered Grading; Safety Management Systems

1. Introduction

Pilot training produces high-stakes human performance outcomes. A student may pass written examinations, accumulate required hours, and complete a syllabus, yet still carry unresolved gaps in radio discipline, checklist discipline, approach stabilization, emergency decision-making, weather judgment, IFR procedural accuracy, or multi-task management. Traditional training records often document these weaknesses after a failed stage check or practical

* Corresponding author: Munashe Naphtali Mupa.

test. Predictive training analytics changes the timing: it seeks to identify risk signals while there is still enough time for targeted instruction, scenario-based practice, and chief-instructor review.

The FAA Airman Certification Standards connect knowledge, risk management, and skill to the privileges of the certificate or rating being exercised (FAA, 2024a; FAA, 2024b). FAA training guidance also emphasizes scenario-based instruction, higher-order thinking skills, learner-centered grading, and structured aeronautical decision-making using models such as PAVE, 3P, 5P, CARE, TEAM, and DECIDE (FAA, 2022; FAA, 2024c; FAA, 2024d). Part 141 schools add a formal curricular architecture, including approved training course outlines, stage checks, end-of-course tests, chief-instructor oversight, and records, while Part 61 instruction has greater flexibility but can still adopt data-driven competency gates (eCFR, 2026a; eCFR, 2026b).

This paper argues that pilot training institutions should treat lesson-stage records as a safety-relevant data asset. The same logic used in enterprise risk management, internal controls, compliance analytics, and interpretable machine learning can be adapted to flight instruction, provided the design is transparent, non-punitive, instructor-governed, and aligned with safety management principles. Mupa and co-authors have emphasized the role of artificial intelligence, risk scoring, internal controls, and analytics in strengthening risk-management systems; those concepts are translated here into an aviation training context where the operational objective is not automation for its own sake, but earlier recognition of learning gaps that can be remediated before they become safety-critical patterns (Mupa et al., 2025a; Mupa et al., 2025b; Netshifhefhe et al., 2024).

1.1. Research Questions

- Which student-training variables are most informative for early detection of checkride risk and safety-relevant learning gaps?
- Can machine-learning models classify candidates into practical risk tiers with sufficient accuracy for instructional intervention?
- How can analytics outputs be translated into non-punitive competency-based training actions under Part 61 and Part 141 operating structures?
- Which governance controls are necessary to prevent predictive training analytics from becoming unfair, punitive, or over-reliant on opaque model outputs?

2. Regulatory and Literature Foundation

The ACS and PTS framework matters because practical tests are not generic flight evaluations. They are structured assessments of defined areas of operation and tasks. The Federal Register confirmed the regulatory incorporation of the ACS and PTS as testing standards, explaining that applicants must complete specific tasks and maneuvers to prescribed standards and that unsatisfactory performance can result in notice of disapproval or denial (Federal Register, 2024). Predictive training analytics should therefore be mapped to ACS task groups rather than to vague labels such as “good student” or “weak student.”

The aviation-instruction literature also supports the need for scenario-based and learner-centered methods. FAA instructor guidance describes scenario-based training as a method that places learners into realistic operational settings and supports decision-based objectives rather than rote maneuver completion (FAA, 2024c; FAA, 2024d). Learner-centered grading and collaborative assessment are particularly relevant because predictive analytics should not replace instructor judgment; it should structure the debrief so that both instructor and learner can see where repeated behaviors indicate a deeper competency gap.

Public aviation safety data further supports the selection of risk variables. The NTSB provides downloadable civil aviation accident data from 1962 onward, with updated public datasets and metadata (NTSB, 2026). Kaggle mirrors of the NTSB aviation accident database provide accessible structured files for aviation safety analysis (Eleraky, 2024). NASA’s Aviation Safety Reporting System provides voluntary and confidential safety-report narratives that help identify recurring operational threats, including communication, procedure, weather, automation, and workload issues (NASA ASRS, 2026). These public sources do not provide student lesson records, but they provide credible risk themes for constructing training analytics variables.

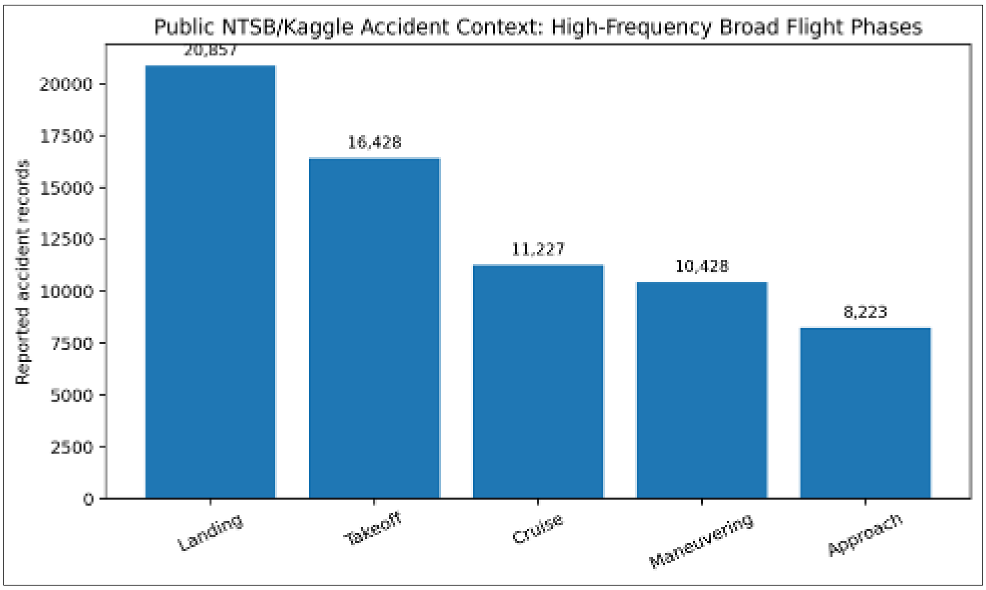

Figure 1 Public accident-context distribution by broad flight phase, used only to anchor training-risk domains and not as a substitute for student records

3. Data and Methodology

The study uses a two-layer data design. The first layer is public-source contextual aviation safety evidence from NTSB, Kaggle mirrors of NTSB data, NASA ASRS, FAA ACS materials, FAA risk-management guidance, and Part 61/Part 141 rules. The second layer is a reproducible synthetic institutional training dataset designed to resemble the structure of records that a flight school could lawfully and ethically collect. This approach is necessary because personally identifiable flight-training records, stage-check outcomes, and practical-test readiness records are not publicly available at scale.

Table 1 Data sources and analytical role

Source layer	Nature of data	Use in the paper
NTSB aviation accident data	Public accident and selected incident records from 1962 to present; downloadable datasets and metadata.	Used to anchor broad safety-risk domains such as phase of flight, weather, and event context.
Kaggle NTSB aviation accident dataset	Structured CSV mirror of NTSB accident records, accessible for reproducible exploratory analysis.	Used to support aviation safety analytics demonstration and figure context.
FAA ACS / PTS materials	Task-level knowledge, risk-management and skill standards for certificates and ratings.	Used to map training variables to practical-test competencies.
FAA Risk Management and Aviation Instructor guidance	ADM, scenario-based training, learner-centered grading, risk assessment and mitigation concepts.	Used to design intervention rules and debrief methodology.
NASA ASRS	Voluntary confidential safety reports from frontline aviation personnel.	Used as qualitative context for non-punitive reporting and recurring safety themes.
Synthetic training-record dataset	2,400 simulated records representing lesson-stage variables a flight school could collect.	Used to demonstrate predictive modeling without exposing real student records.

The synthetic dataset contains 2,400 candidate-stage records across five training tracks: Private Pilot, Instrument Rating, Commercial Pilot, CFI/CFII, and Multi-Engine Add-on. Each record represents a student approaching a stage-check or practical-test readiness decision. The binary outcome is a modeled "checkride risk outcome," meaning that the candidate displays a pattern consistent with elevated risk of practical-test non-approval or safety-relevant readiness deficiency. It is not a real FAA practical-test result and should not be read as a national pass/fail statistic.

Table 2 Candidate predictors for checkride-risk detection

Variable	Type	Rationale
Repeat lessons	Count	Number of lessons repeated before stage readiness; proxy for persistence of unresolved deficiencies.
Unsatisfactory maneuvers	Count	Instructor-coded maneuver outcomes below standard during stage preparation.
Procedural deviation rate	Continuous	Frequency of checklist, profile, callout, altitude, speed, or task-sequencing deviations.
Radio readback error percent	Continuous	Communication accuracy signal relevant to IFR, ATC-density, and runway safety.
IFR tasks incomplete	Count	Holds, approaches, missed approaches, clearance compliance, or partial-panel tasks not yet meeting standard.
Approach-stability violations	Count	Unstable approach criteria or repeated correction patterns.
Weather risk exposure	1-5 scale	Training exposure to MVFR/IMC/night/crosswind conditions; contextual risk exposure only.
Workload density score	1-5 scale	Airspace, ATC, procedure, frequency-change, or task saturation intensity.
Instructor interventions	Count	Instructor-required corrections, safety calls, or task resets.
Safety event flag	Binary	Non-punitive flag for runway incursion risk, loss of separation concern, checklist omission, or unstable continuation.
Knowledge test score	Percent	Ground knowledge strength; included because weak knowledge can predict practical-test risk.
Stage-check score	Percent	Pre-practical competency score reflecting ACS-aligned performance.
Lesson gap days	Continuous	Time gap between training events, used as a continuity/skill-decay proxy.
Instructor risk terms	Count	Structured count of debrief terms such as "behind aircraft," "unstable," "missed item," or "late correction."
Checklist adherence score	Percent	Composite checklist discipline measure.

3.1. Descriptive Analytics

The descriptive layer shows the type of dashboard a chief instructor or training manager would need before using predictive models. The simulated risk outcome rate is 12.4%, with higher modeled risk in tracks involving greater procedural complexity, teaching standards, IFR exposure, or asymmetric-thrust decision-making. These figures are not national FAA pass/fail rates; they are a reproducible modeling environment used to test whether training variables can be converted into actionable risk tiers.

Table 3 Synthetic training-record descriptive summary

Metric	Value
Synthetic cases	2,400
Overall risk outcome rate	12.4%
Mean stage check score	76.3
Mean knowledge test score	82.0
Mean procedural deviation rate	11.9
Mean repeat lessons	1.49
Mean instructor interventions	2.67
Mean safety event flag	9.5%

Table 4 Modeled training risk by track

Training track	Cases	Risk outcome rate (%)	Avg stage check score	Avg procedural deviation rate	Safety event rate (%)
CFI/CFII	373	20.9	75.2	12.8	13.7
Commercial Pilot	392	5.9	77.6	11.2	6.6
Instrument Rating	681	13.5	75.7	12.1	10.4
Multi-Engine Add-on	216	34.7	73.0	14.0	13.4
Private Pilot	738	3.9	77.6	10.9	7.0

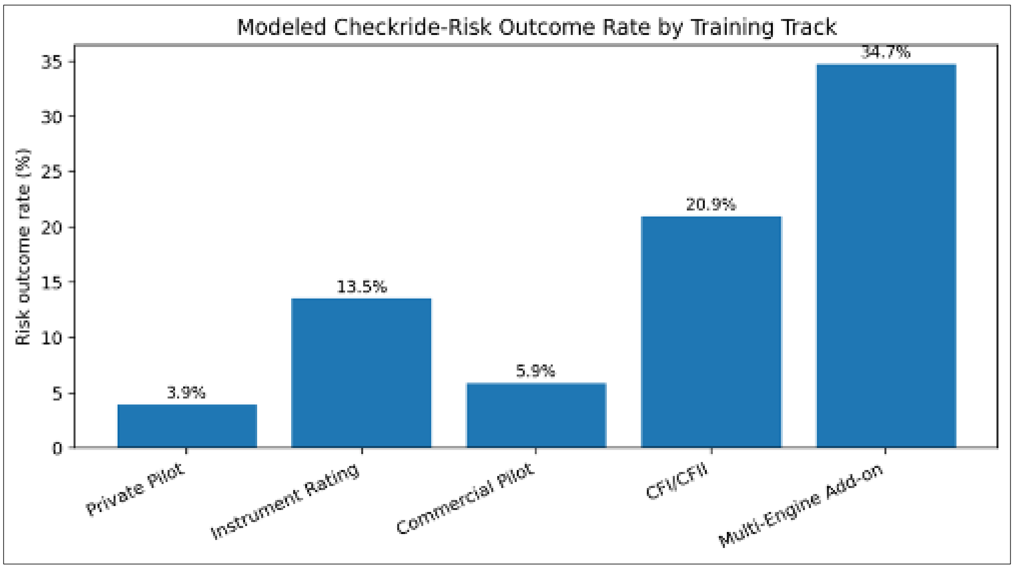

Figure 2 Modeled checkride-risk outcome rate by training track in the reproducible simulation dataset.

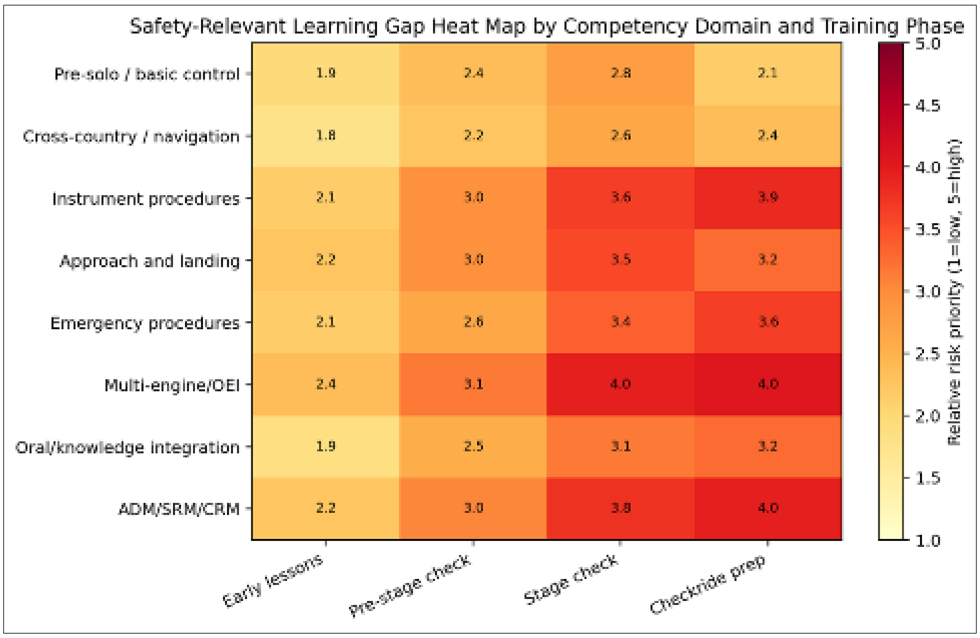

Figure 3 Competency-domain heat map showing where safety-relevant learning gaps intensify across the training sequence

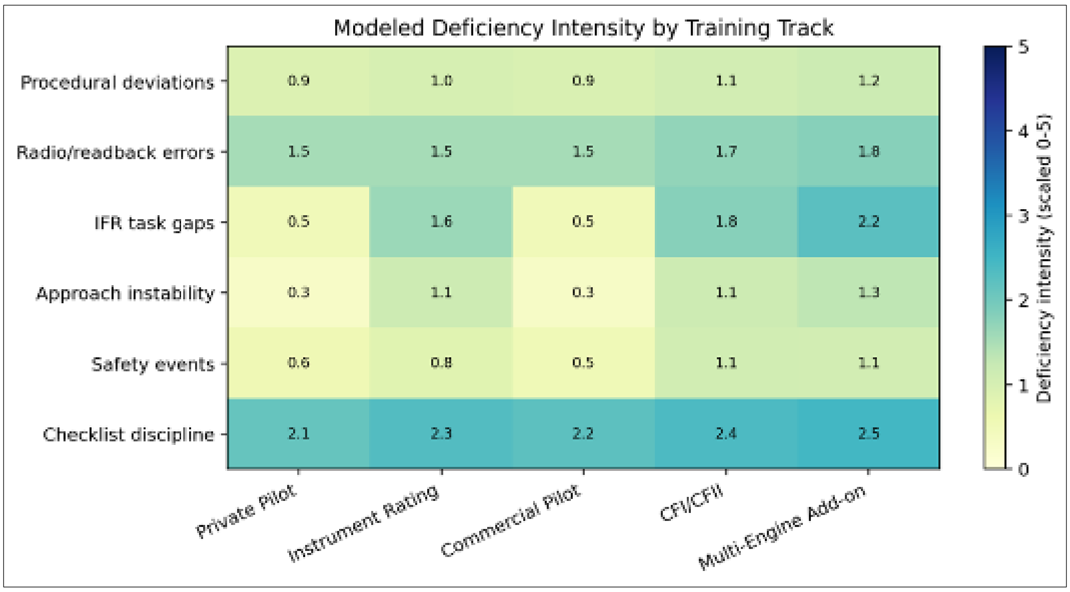

Figure 4 Deficiency-intensity heat map by track, useful for chief-instructor curriculum review and targeted remediation

3.2. Predictive Modeling Design

The modeling design deliberately compares interpretable and non-linear algorithms. Logistic regression provides transparent directionality and approximate odds interpretation after standardization. Random forests and gradient boosting can capture non-linear interactions, such as the combined effect of low stage-check scores, repeated unsatisfactory maneuvers, high workload exposure, and instructor interventions. The models were trained on 70% of the dataset and tested on 30%, preserving the same risk-outcome proportion across train and test sets. Five-fold cross-validation AUC was also calculated to reduce overfitting risk.

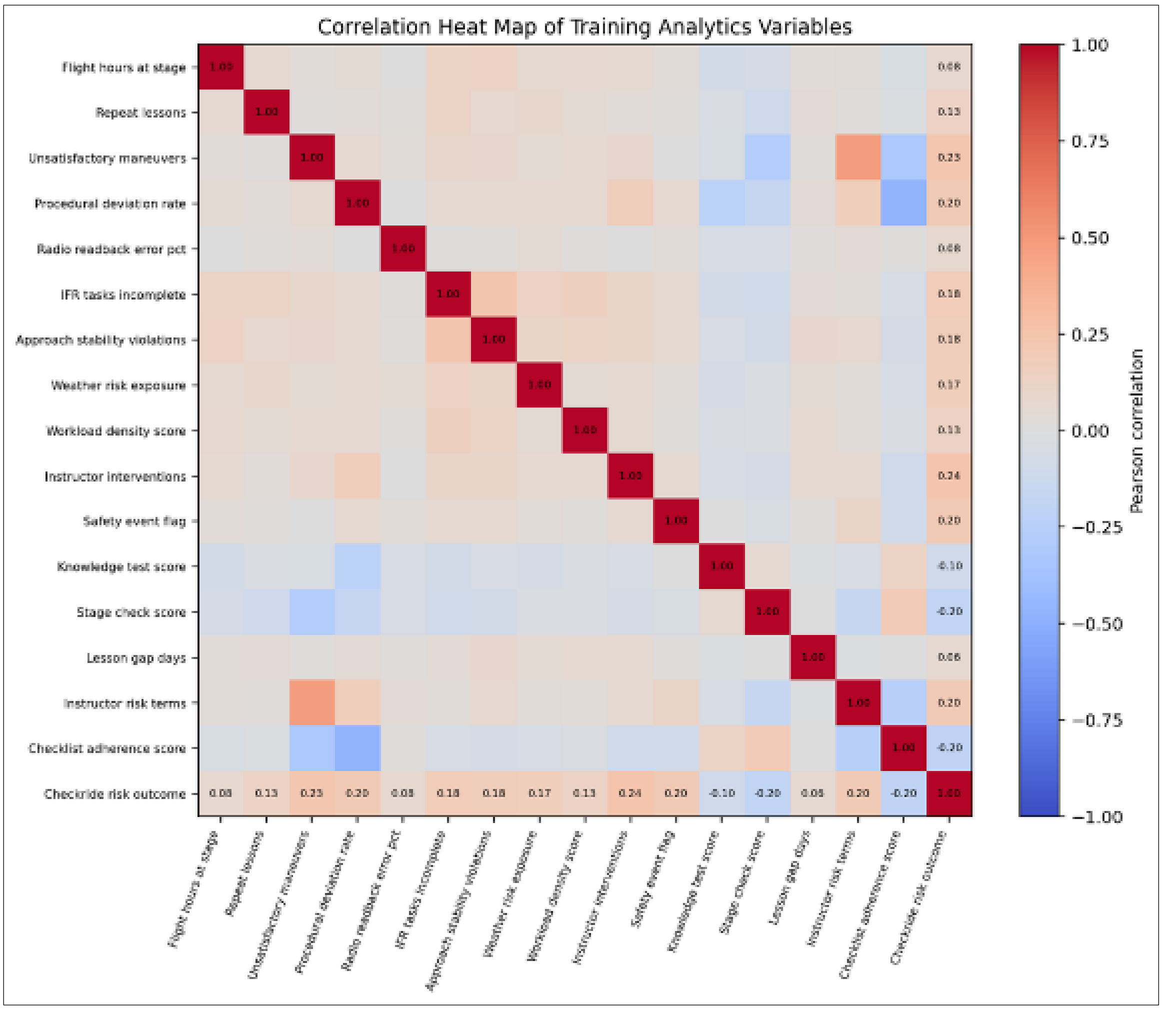

Figure 5 Correlation heat map across candidate predictors and the modeled checkride-risk outcome.

Table 5 Model performance comparison

Model	Test AUC	5-fold CV AUC	Accuracy	Precision	Recall	F1	Brier
Logistic regression	0.837	0.850	0.750	0.298	0.753	0.427	0.160
Random forest	0.818	0.833	0.860	0.430	0.416	0.423	0.110
Gradient boosting	0.793	0.831	0.868	0.421	0.180	0.252	0.095

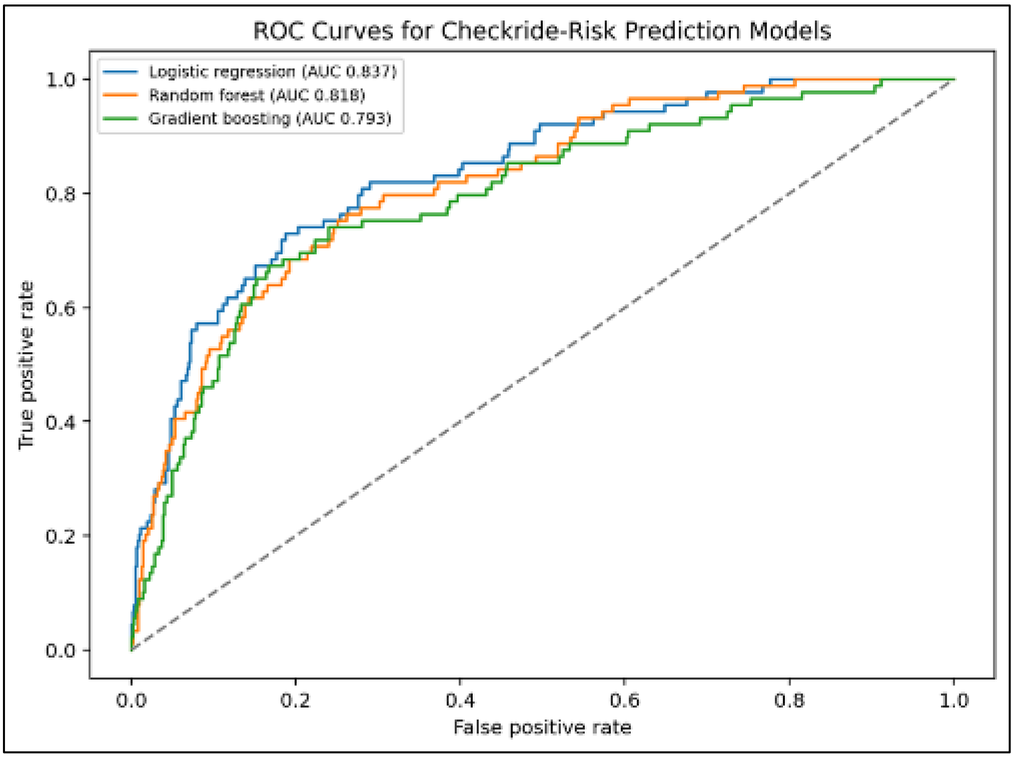

Figure 6 ROC curves for model comparison. Logistic regression achieved the strongest discriminative performance in this simulation.

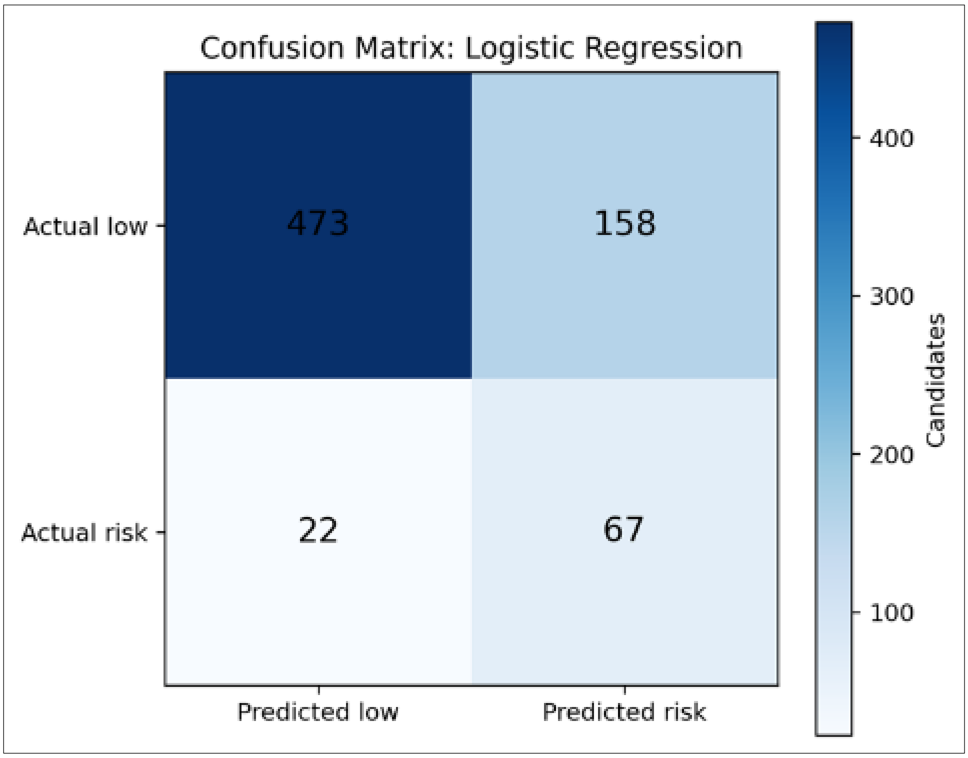

Figure 7 Confusion matrix for the selected logistic-regression model at a 0.50 threshold.

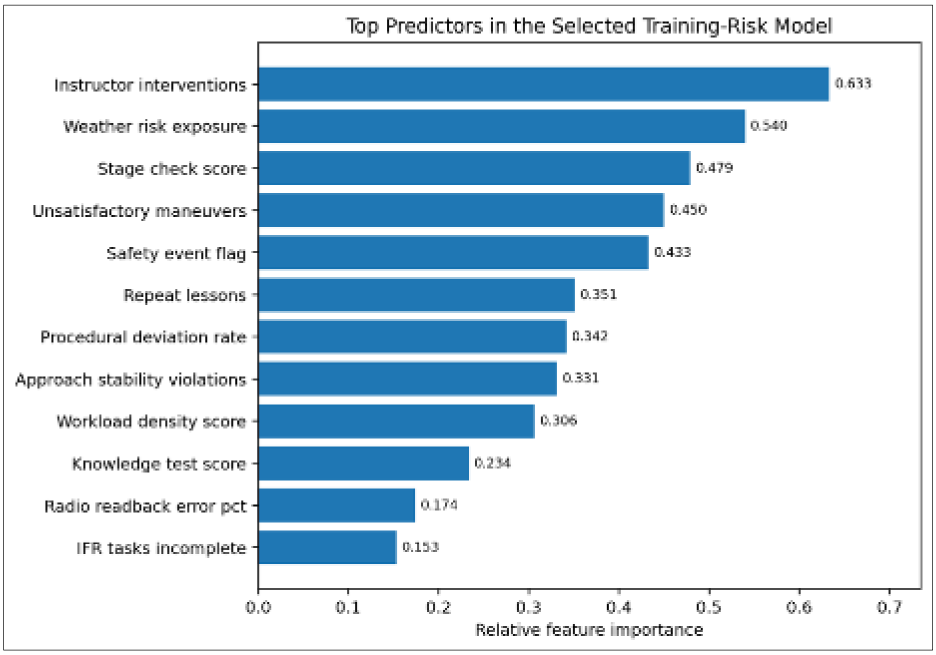

Figure 8 Feature-importance ranking for the selected model. High importance should trigger review, not automatic punitive action

4. Results and Interpretation

The results support the proposition that early training records can be converted into useful practical-test readiness indicators. In the simulation, the strongest predictors were not merely total flight hours; they were quality and recurrence indicators: stage-check score, checklist adherence, instructor interventions, unsatisfactory maneuvers, procedural deviation rate, IFR task gaps, approach stability, safety-event flags, lesson continuity, and workload exposure. This is important because an hour-based training culture can hide repeated errors. Competency-based analytics model instead asks whether the learner repeatedly meets, misses, or recovers to standard under realistic workload conditions.

Table 6 Top 10 predictors in the selected model

Predictor	Relative importance
Instructor interventions	0.633
Weather risk exposure	0.540
Stage check score	0.479
Unsatisfactory maneuvers	0.450
Safety event flag	0.433
Repeat lessons	0.351
Procedural deviation rate	0.342
Approach stability violations	0.331
Workload density score	0.306
Knowledge test score	0.234

Table 7 Directional interpretation from standardized logistic regression

Predictor	Standardized logistic coefficient
Instructor interventions	0.633
Weather risk exposure	0.540
Stage check score	-0.479
Unsatisfactory maneuvers	0.450
Safety event flag	0.433
Repeat lessons	0.351
Procedural deviation rate	0.342
Approach stability violations	0.331

The model output should be interpreted as a structured warning, not a verdict. For example, a candidate assigned to the "High" tier should receive a chief-instructor review, a documented debrief, focused scenario-based retraining, and a fresh readiness assessment. The model should not be used to deny training access, punish students, or pressure instructors to withhold endorsements without an individualized review. This is consistent with a safety management approach that treats weak signals as opportunities to correct system and training design, not as grounds for blame.

Table 8 Risk-tier calibration in the test sample

Risk tier	Candidates	Actual_risk_rate	Avg_predicted_risk
Low	277	2.2	6.4
Monitor	119	7.6	21.8
Elevated	99	7.1	39.4
High	225	29.8	72.8

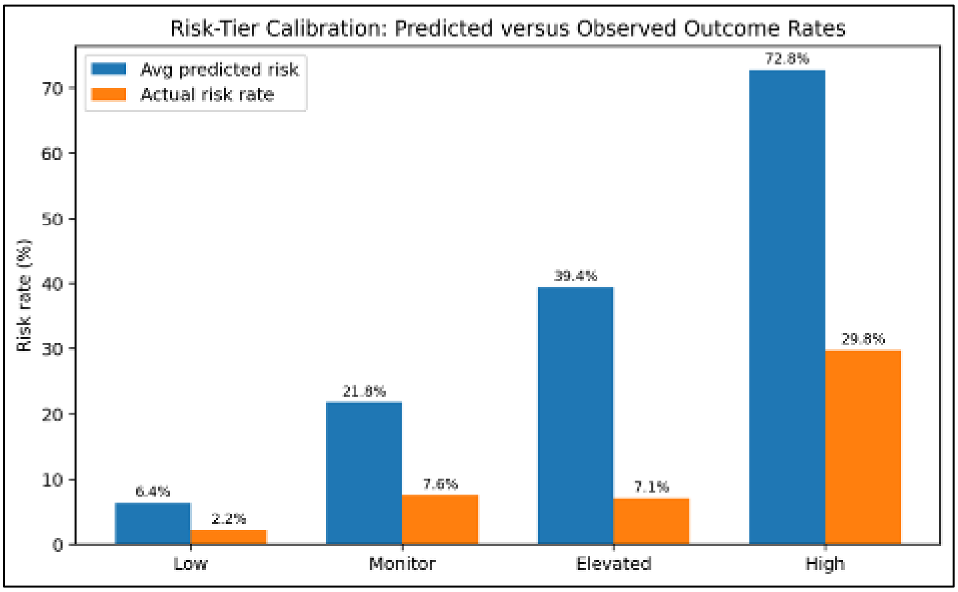

Figure 9 Risk-tier calibration comparing average predicted risk against actual modeled risk outcomes.

4.1. Proposed Predictive Training Analytics Framework

The proposed framework has six components: data governance, ACS-coded variables, predictive modeling, instructor review, non-punitive remediation, and outcome monitoring. The chief instructor remains the instructional authority. The model supplies pattern recognition: it can flag repeated deviations, combinations of weaknesses, or changes in performance trajectory that may be difficult to see from isolated lesson notes.

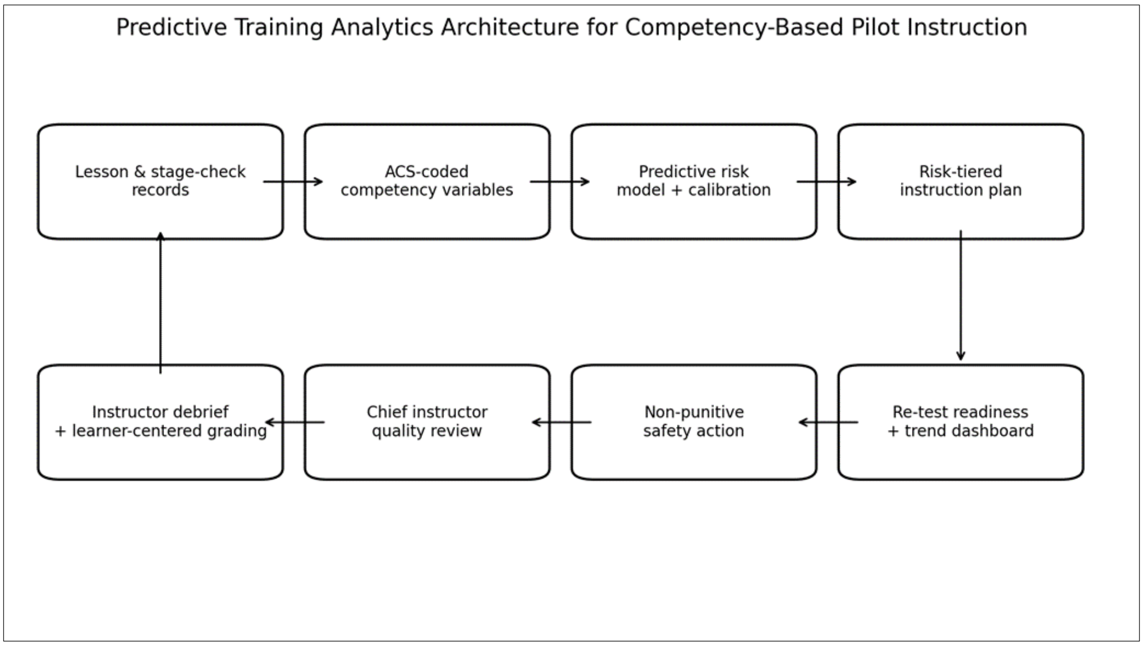

Figure 10 Proposed architecture for predictive training analytics in competency-based pilot instruction.

Table 9 Predictive training analytics operating framework

Component	Operational content	Governance purpose
Data capture	Lesson records, simulator records, stage checks, instructor comments, knowledge scores, safety flags.	Standardize definitions and avoid free-text-only records.
ACS coding	Map each deficiency to ACS/PTS task, risk-management element, skill element, or knowledge area.	Prevents vague judgments and improves fairness.
Risk scoring	Use interpretable models and calibrated probabilities to classify risk tiers.	Support early intervention without replacing instructor judgment.
Instructional intervention	Assign scenario-based remediation, chair flying, simulator practice, briefing/debrief gates, or chief-instructor flights.	Convert analytics into training action.
Quality assurance	Review false positives, false negatives, instructor variance, and cohort differences.	Prevent bias, drift, and overreliance.
Safety feedback loop	Track stage-check repeat rates, practical-test readiness, safety events, and recurring deficiency themes.	Improve syllabus design and instructor standardization.

Table 10 Risk-tiered instructional response plan

Risk tier	Probability band	Training response	Governance control
Low	<15%	Continue normal syllabus; maintain standard debrief; monitor only.	No special action.
Monitor	15-30%	Add targeted homework, chair flying, and next-lesson proficiency check on flagged domain.	Instructor documents response and trend.
Elevated	30-50%	Require focused scenario-based remediation, simulator or ground review, and a supervised readiness review.	Chief instructor reviews before endorsement or stage progression.
High	>50%	Pause checkride endorsement; conduct root-cause review, safety debrief, and individualized remediation plan.	Non-punitive safety review and documented re-assessment before release.

4.2. Ethical, Fairness, and Safety-Management Controls

Predictive analytics in flight training must be governed carefully. A model that is technically accurate can still be harmful if it is used to stigmatize learners, pressure instructors, ignore context, or create hidden barriers to progression. The framework therefore requires fairness auditing, transparency, appealability, data minimization, instructor override, and continuous calibration. The objective is early support and safety assurance, not algorithmic gatekeeping.

- Non-punitive use: Flags trigger review and support, not automatic failure or denial of endorsement.
- Human-in-the-loop review: The chief instructor or designated standardization instructor must review high-risk cases before action.
- Explainability: Each risk alert must show the top contributing variables and the relevant ACS task domain.
- Privacy and minimization: Training records should be limited to instructionally necessary data, protected from unnecessary disclosure, and retained under school policy.
- Fairness audit: Outcomes should be reviewed by instructor, aircraft type, training track, lesson gaps, and cohort to detect bias or inconsistent grading.
- Model drift monitoring: Feature importance, false positives, false negatives, and calibration should be re-tested every training cycle.

Table 11 Governance controls for predictive training analytics

Risk	Control	Review cycle
Bias from instructor scoring differences	Standardization meetings, rubric calibration flights, inter-rater review.	Monthly
Overreliance on model outputs	Mandatory human review and documented instructor judgment.	Every flagged case
False positives	Debrief-only first response; no automatic punitive consequence.	Every elevated/high case
False negatives	Review failures and safety events that were not flagged; update variables.	Quarterly
Privacy leakage	Role-based access, de-identification for dashboards, retention policy.	Continuous
Model drift	Recalculate AUC, Brier score, calibration, feature importance.	Each semester or 500 records

4.3. Implementation Roadmap for Part 61 and Part 141 Environments

Implementation should be phased. A small Part 61 school can begin with spreadsheet-based ACS coding and debrief rubrics. A Part 141 school can integrate the framework with stage-check records, chief-instructor oversight, and standardization meetings. In both environments, the first objective is better instructional visibility, not immediate automation.

Table 12 Implementation roadmap

Phase	Timeline	Activities	Output
Phase 1: Design	0-30 days	Define ACS-coded variables, data dictionary, privacy controls, risk-tier rules, and instructor rubric.	Approved data dictionary and debrief form.
Phase 2: Pilot capture	31-90 days	Collect lesson-stage data without predictive consequences; review data completeness and instructor consistency.	Data-quality report and rubric calibration.
Phase 3: Model validation	91-150 days	Train initial model, test AUC/calibration, review false positives/negatives, and adjust thresholds.	Validation memorandum and threshold policy.
Phase 4: Supervised use	151-240 days	Use risk alerts for debrief support and chief-instructor review only.	Intervention logs and student outcome tracking.
Phase 5: Continuous improvement	After 240 days	Monitor drift, fairness, performance, instructor variance, and syllabus changes.	Quarterly safety-training analytics report.

5. Discussion

The principal contribution of this paper is to move checkride-readiness analysis from subjective end-point judgment toward structured early-warning design. The analysis shows that risk is multi-dimensional. A learner may have acceptable hours but still display repeated procedural deviations, radio-readback errors, unstable-approach patterns, high instructor intervention counts, or repeated IFR task incompletion. Conversely, a learner with one poor lesson should not be over-classified as high risk if the broader trend shows recovery, strong checklist adherence, and improving stage-check performance.

The paper also contributes to flight-training standardization. Part 141 schools often possess records, stage checks, and standardization structures, but the records may remain descriptive rather than predictive. Part 61 operations may be more flexible, but they can still adopt light-weight analytics by using common debrief codes, ACS task mappings, and repeated-deficiency dashboards. The central idea is that a training record should not merely prove that hours were flown; it should show how competency, risk management, and decision quality developed over time.

The Mupa risk-management literature is relevant because aviation training analytics is fundamentally a controls problem: define risk, capture evidence, assign responsibility, detect anomalies, intervene early, document response, and review outcomes. The same principles used in AI-enabled risk management, internal controls, audit planning, and digital transformation can support flight-training governance when adapted to aviation's safety culture and human-performance realities (Mupa et al., 2025a; Mupa et al., 2025b; Netshifhefhe et al., 2024; Mupa et al., 2024).

Limitations

- The student-training dataset is synthetic. It demonstrates method and architecture but does not represent real student records, real school outcomes, or official FAA pass/fail results.
- Public NTSB and ASRS sources are safety-context sources, not direct checkride datasets. They support variable selection and risk-domain framing but cannot prove individual student readiness.
- The model does not include examiner variance, aircraft-specific factors, weather on test day, student stress, financial interruptions, or instructor-student relationship factors unless a school captures them ethically.
- AUC and feature importance in a simulation may overstate performance compared with real deployment. Real-world implementation would require prospective validation and recurrent calibration.

- The framework must be adapted to local syllabus design, aircraft fleet, instructor standardization maturity, and data governance capability.
- The paper does not recommend replacing flight instructors, designated pilot examiners, ACS standards, or chief-instructor judgment with machine learning.

6. Conclusion

Predictive training analytics can strengthen competency-based pilot instruction by shifting attention from late failure to early remediation. When lesson records are standardized around ACS task domains, procedural discipline, risk management, and instructor interventions, they can reveal patterns that are difficult to see in isolated debriefs. The selected logistic-regression model performed well in a reproducible synthetic demonstration, but the deeper value of the framework is not the algorithm itself. The value lies in the operating model: structured data capture, non-punitive risk-tiering, learner-centered remediation, chief-instructor quality review, and continuous monitoring of training outcomes.

The recommended pathway is incremental. Flight schools should begin with a data dictionary, standardized debrief codes, ACS task mapping, and simple dashboards. Predictive modeling should be introduced only after the data are consistent and governance controls are in place. Used properly, the framework can help instructors detect checkride risk, procedural noncompliance, and safety-relevant learning gaps earlier, while preserving the professional judgment and safety culture that remain central to pilot training.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Breiman, L. (2001) “Random forests”, Machine Learning, 45(1), pp. 5-32.

[2] eCFR (2026a) 14 CFR Part 61 - Certification: Pilots, Flight Instructors, and Ground Instructors. Electronic Code of Federal Regulations. Available at: https://www.ecfr.gov/current/title-14/chapter-I/subchapter-D/part-61 (Accessed: 17 June 2026).

[3] eCFR (2026b) 14 CFR Part 141 - Pilot Schools. Electronic Code of Federal Regulations. Available at: https://www.ecfr.gov/current/title-14/chapter-I/subchapter-H/part-141 (Accessed: 17 June 2026).

[4] Eleraky, Y. (2024) Aviation Accident NTSB. Kaggle dataset. Available at: https://www.kaggle.com/datasets/yassereleraky/aviation-accident-ntsb (Accessed: 17 June 2026).

[5] FAA (2022) Risk Management Handbook, FAA-H-8083-2A. Washington, DC: Federal Aviation Administration.

[6] FAA (2024a) Instrument Rating - Airplane Airman Certification Standards, FAA-S-ACS-8C. Washington, DC: Federal Aviation Administration.

[7] FAA (2024b) Commercial Pilot for Airplane Category Airman Certification Standards, FAA-S-ACS-7B. Washington, DC: Federal Aviation Administration.

[8] FAA (2024c) Aviation Instructor’s Handbook: The Learning Process. Washington, DC: Federal Aviation Administration.

[9] FAA (2024d) Aviation Instructor’s Handbook: Assessment. Washington, DC: Federal Aviation Administration.

[10] FAA (2026) U.S. Civil Airmen Statistics. Washington, DC: Federal Aviation Administration. Available at: https://www.faa.gov/data_research/aviation_data_statistics/civil_airmen_statistics (Accessed: 17 June 2026).

[11] Federal Register (2024) Airman Certification Standards and Practical Test Standards for Airmen; Incorporation by Reference. Federal Register, 1 April. Available at: https://www.federalregister.gov/documents/2024/04/01/2024-06644/airman-certification-standards-and-practical-test-standards-for-airmen-incorporation-by-reference (Accessed: 17 June 2026).

[12] Friedman, J. H. (2001) “Greedy function approximation: A gradient boosting machine”, Annals of Statistics, 29(5), pp. 1189-1232.

[13] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn. New York: Springer.

[14] Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X. (2013) Applied Logistic Regression. 3rd edn. Hoboken: Wiley.

[15] Mupa, M. N. et al. (2024) “The evolving role of management accountants in risk management and internal controls in the energy sector”, available via ResearchGate profile: Munashe Naphtali Mupa.

[16] Mupa, M. N. et al. (2025a) “Risk and compliance paper: what role does Artificial Intelligence (AI) play in enhancing risk management practices in corporations”, World Journal of Advanced Research and Reviews, 27(1). DOI: 10.30574/wjarr.2025.27.1.2607.

[17] Mupa, M. N. et al. (2025b) “Interpretable machine learning for audit planning: Improving misstatement and compliance risk detection in financial services”, World Journal of Advanced Research and Reviews, 28(2), pp. 925-933. DOI: 10.30574/wjarr.2025.28.2.3779.

[18] NASA ASRS (2026) Aviation Safety Reporting System Database Online. National Aeronautics and Space Administration. Available at: https://asrs.arc.nasa.gov/search/database.html (Accessed: 17 June 2026).

[19] Netshifhefhe, K., Netshifhefhe, M. V., Mupa, M. N. and Mupa, K. A. M. (2024) “Integrating internal auditing and legal compliance: A strategic approach to risk management”, IRE Journals. Available via Google Scholar profile: Munashe Naphtali Mupa.

[20] NTSB (2026) Accident Data. National Transportation Safety Board. Available at: https://www.ntsb.gov/safety/data/Pages/Data_Stats.aspx (Accessed: 17 June 2026).

[21] Molnar, C. (2022) Interpretable Machine Learning. 2nd edn. Available at: https://christophm.github.io/interpretable-ml-book/ (Accessed: 17 June 2026).

[22] Wickens, C. D., Lee, J., Liu, Y. and Gordon-Becker, S. (2004) An Introduction to Human Factors Engineering. 2nd edn. Upper Saddle River: Pearson.

Appendix A. Synthetic Data Generating Process

The reproducible simulation was generated with a fixed random seed. The target outcome was generated through a logistic risk function combining repeat lessons, unsatisfactory maneuvers, procedural-deviation rate, radio-readback errors, IFR task incompletion, approach-stability violations, weather exposure, workload density, instructor interventions, safety-event flag, knowledge-test score, stage-check score, lesson gaps, instructor-comment risk terms, checklist adherence, and track complexity. The model intentionally makes hours less predictive than competency quality, consistent with the paper’s argument that competency-based readiness requires more than time accumulation.

Table A1 Modeling specifications

Model component	Specification
Train/test split	70/30 stratified split
Outcome type	Binary modeled checkride-risk outcome
Selected model	Logistic regression classifier
Threshold	0.50 for confusion matrix; tier thresholds used for intervention
Primary metric	ROC AUC
Secondary metrics	Accuracy, precision, recall, F1, Brier score, calibration table
Governance rule	No automatic punitive action; all elevated/high flags require human review