Volume 4 / Issue 8/ August 2026 International Journal of Modern Science and Research Technology
ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

Artificial Intelligence Enabled Human Factors Risk Prediction for Reducing Aviation Accidents in Highly Automated Flight Operations

Idoko Peter Idoko
Department of System Safety and Reliability Studies, Centre for Safety Management, Federal University Lokoja, Nigeria

Victoria Bukky Ayoola
Department of Climate Change Resilience and Sustainability, Federal University Lokoja, Nigeria

Abstract—The growing use of flight-management systems, autopilot, autothrust, and other advanced automation has shifted flightcrew work from continuous manual control toward supervisory monitoring and exception management. This study develops and evaluates an Artificial Intelligence Enabled Human Factors Risk Prediction (AI-HFRP) framework that integrates pilot state, human-automation interaction, flight dynamics, and operational context into an interpretable safety-risk estimate. The supplied proof-of-concept dataset contained 24 realistic synthetic pilot-flight windows spanning four flight profiles and six phases of flight. Human-state inputs included Samn-Perelli fatigue, Karolinska Sleepiness Scale, NASA Task Load Index mental demand, Psychomotor Vigilance Task response measures, and Situation Awareness Global Assessment Technique accuracy. Automation and operational inputs included mode transitions, mode awareness, mode confusion, automation surprise, Flight Management System activity, cockpit alerts, flight-path deviations, visibility, crosswind, and air-traffic-control activity. High and Critical expert labels were combined for binary evaluation because only five such observations were available. Flight-grouped four-fold cross-validation was used to limit leakage between observations from the same flight. The proposed interpretable fusion achieved an area under the receiver operating characteristic curve of 0.874, sensitivity of 0.800, specificity of 0.684, and F1-score of 0.533. Benchmark model AUC values ranged from 0.547 to 0.695. Elevated risk was associated with longer duty, greater fatigue and sleepiness, slower vigilance response, reduced situation awareness, automation surprise, and larger flight-path deviations. The results support further evaluation of human-centred predictive safety assistance, but larger operational datasets are required before deployment or accident-reduction claims are justified.

Keywords:
aviation safety, artificial intelligence, human factors, cockpit automation, pilot fatigue, situation awareness, machine learning, predictive safety.

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

I. Introduction

A. Background of the Study

Modern commercial aircraft increasingly rely on integrated autopilot, autothrust, Flight Management Systems (FMS), Flight Directors, electronic alerting systems, and flight-envelope functions to improve flight-path precision, efficiency, and operational consistency. This technological development has progressively shifted flightcrew activity from continuous manual control toward supervisory monitoring, system management, and exception handling. The safety benefit of automation is substantial, but the change in work structure creates a human-factors requirement: pilots must continuously understand the active automation state, anticipate future mode behavior, and retain sufficient situation awareness to intervene when the automated system no longer behaves as expected.

Established human-factors instruments provide measurable indicators of pilot cognitive and behavioral state. NASA Task Load Index (NASA-TLX) measures multidimensional workload [1]; the Samn-Perelli scale supports repeated aircrew fatigue assessment [2]; the Karolinska Sleepiness Scale (KSS) measures momentary sleepiness [3], [4]; Situation Awareness Global Assessment Technique (SAGAT) evaluates situation awareness in dynamic simulations [5], while the Situation Awareness Rating Technique (SART) provides a complementary subjective assessment [6]. Psychomotor Vigilance Task (PVT) measures are sensitive to fatigue-related performance degradation [7], and validated trust-in-automation instruments provide a structured means of examining whether reliance on automated systems remains appropriately calibrated [8]. Synchronizing these measures with objective automation and flight-state data provides a stronger basis for multimodal predictive modelling than using an arbitrary composite human-error score.

The Federal Aviation Administration Human Factors Team identified automation awareness, flight-path monitoring, mode interpretation, and unexpected automation behavior as important interface issues in modern flight decks [9]. Highly automated operations can therefore generate vulnerabilities involving mode confusion, automation surprise, overreliance, vigilance decrement, workload transitions, and delayed manual takeover. These vulnerabilities become more consequential when combined with fatigue, reduced sleep, weather, late air-traffic-control clearances, high alert burden, or system abnormalities.

Consistent with current human-centred aviation AI guidance, the present study treats AI as a predictive safety assistant rather than a replacement for pilot authority [10]. Objective and subjective evidence can also be combined across complementary sources: pilot-workload research has paired NASA-TLX with heart-rate-variability measures [11]; standardized occurrence and flight-phase coding can be aligned with ICAO ADREP taxonomy [12]; FAA Flight Operational Quality Assurance (FOQA) provides routine digital flight-data evidence [13]; NASA Aviation Safety Reporting System (ASRS) reports provide de-identified human-factors narratives [14]; and National Transportation Safety Board investigations provide detailed accident and selected-incident evidence [15]. These resources remain essential but are largely retrospective or event-triggered, creating a strong rationale for predictive methods capable of detecting converging precursor patterns before safety margins are substantially reduced.

B. Problem Statement

Human-performance degradation in highly automated flight can develop gradually and remain difficult to recognize until task demand increases. Fatigue, workload imbalance, reduced vigilance, loss of situation awareness, mode confusion, and automation surprise may each appear modest in isolation but become safety-relevant when they occur together with adverse weather, repeated mode changes, cockpit alerts, or flight-path deviations. Existing safety analyses also tend to separate human-state data from aircraft automation and operational context, limiting the ability to recognize combined risk patterns.

A second methodological problem is that aviation machine-learning studies can appear highly accurate when repeated observations from the same pilot or flight are divided randomly between training and testing sets. Post-event variables can also leak information about the target. A credible predictive model therefore requires explicit separation of predictors and outcomes, flight-grouped validation, and interpretable outputs that can be audited by aviation professionals.

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

C. Aim and Objectives of the Study

The aim is to develop and conduct a proof-of-concept evaluation of an AI-enabled framework for predicting human-factors-related risk in highly automated flight operations. The objectives are to identify influential human and automation indicators; examine relationships among pilot state, operational conditions, and expert-adjudicated risk; develop an interpretable human-state, automation-interaction, and operational-context fusion model; compare the proposed model with Logistic Regression, Random Forest, Support Vector Machine, Extreme Gradient Boosting, and Artificial Neural Network classifiers; and determine how predictive intelligence could support proactive safety management.

D. Research Questions

- Which pilot-state and human-automation variables show the strongest relationships with increasing risk?
- How do fatigue, sleepiness, workload, vigilance, and situation awareness vary across risk classes?
- How are mode transitions, mode confusion, automation surprise, and cockpit alerts related to risk?
- How does the proposed AI-HFRP framework perform relative to benchmark classifiers?
- Which phases of flight contain the greatest concentration of High/Critical observations?
- How can predictive human-factors intelligence support safety management without reducing flightcrew authority?

E. Significance and Scope

The study integrates validated human-performance constructs with observable automation and flight-context variables instead of relying on an arbitrary single human-error score. It also treats risk as a changing operational state rather than a characteristic assigned only after an occurrence. The framework is relevant to simulator training, fatigue-risk management, Flight Operational Quality Assurance (FOQA), recurrent training, Safety Management Systems (SMS), and human-centred flight-deck design. The supplied dataset contains 24 realistic synthetic pilot-flight windows organized into four flight profiles and six phases. Accordingly, the study is a methodological proof of concept and does not establish operational accident reduction or regulatory readiness.

II. Literature Review

A. Human Factors in Aviation Safety

Human performance in aviation is shaped by interaction among individual capability, task demand, technology, organizational arrangements, and the operating environment. Wiegmann and Shappell's Human Factors Analysis and Classification System (HFACS) links unsafe acts with preconditions, supervision, and organizational influences [16]. Reason's systems approach similarly explains accidents as the alignment of active failures with latent weaknesses across defensive layers [17]. Together, these perspectives imply that human error should not be modelled as an isolated pilot attribute.

Fatigue is important because it affects sustained attention, reaction time, working memory, and monitoring. Samn-Perelli and KSS provide structured subjective measures [2]-[4], whereas PVT measures provide behavioral evidence of vigilance degradation [7]. Workload can impair information processing when excessive, while prolonged low workload can reduce engagement. NASA-TLX remains a widely used multidimensional workload instrument [1]. Situation awareness is also central to automated operations because pilots must perceive relevant information, understand current system state, and project future behavior [5].

B. Highly Automated Flight Operations

Automation can support information acquisition, analysis, decision selection, and action implementation at different levels [18]. Increased automation at one stage may reduce direct manual workload while increasing monitoring and coordination demands elsewhere. The pilot must therefore understand not only the aircraft state but also the intent and logic of the automated system. Mode confusion occurs when the pilot's mental model differs from the active or armed automation state, whereas automation surprise arises when system behavior departs from expectation [9]. Trust must likewise remain calibrated rather than blindly high or low, which is why validated trust-in-automation

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

measurement is relevant to supervisory flight-deck research [8]. Observable indicators such as flight-mode transitions, correct mode identification, FMS reprogramming, unexpected mode behavior, manual override, and takeover latency are therefore more defensible than a generic automation-level score.

C. Human Factors Models and Risk Frameworks

The SHELL framework emphasizes interfaces among Software, Hardware, Environment, and Liveware. The Swiss Cheese Model describes accidents as the alignment of vulnerabilities across defensive barriers [17], while HFACS provides a structured retrospective classification of human and organizational contributors [16]. Threat and Error Management (TEM) focuses on operational threats, crew errors, and undesired aircraft states. AI does not replace these frameworks; it can provide a computational mechanism for integrating indicators that the frameworks identify conceptually.

D. AI and Machine Learning in Aviation Safety

Supervised machine-learning methods estimate relationships between predictors and known outcomes. Logistic Regression provides an interpretable probabilistic baseline. Random Forest can capture nonlinear interactions through ensembles of decision trees [19]. Support Vector Machines construct separating decision surfaces and can perform effectively in high-dimensional settings [20]. Extreme Gradient Boosting combines sequential trees to minimize prediction error [21], while Artificial Neural Networks offer greater functional flexibility but generally require larger datasets [22]. Explainability is essential in safety-critical applications; methods such as SHAP can quantify predictor contributions [23], while intrinsically interpretable fusion models retain domain-level explanations throughout inference.

E. Existing Prediction Approaches and Research Gap

Subjective workload and fatigue scales are inexpensive and validated but are not continuously available in normal flight. Physiological measures such as heart-rate variability, electrodermal activity, eye tracking, and PERCLOS can provide continuous state information but introduce sensor-quality, privacy, and certification issues. Flight data provide objective aircraft-state information but usually cannot determine why a deviation occurred. Alaimo et al. demonstrated the value of considering NASA-TLX alongside heart-rate-variability measures in pilot workload analysis [11]. ICAO ADREP provides a recognized taxonomy for occurrence and event-phase coding [12]. FAA FOQA supports routine operational flight-data monitoring [13], NASA ASRS provides de-identified human-factors narratives [14], and NTSB investigations provide accident and selected-incident evidence [15]. These sources are complementary, but no single source normally contains synchronized human state, automation state, flight dynamics, and outcome labels.

Recent aviation scholarship further supports the transition from retrospective event analysis toward integrated and predictive safety management. Chukwuemeka and Mupa proposed an integrated human-factors approach for accident reduction in the era of advanced flight technology [24], a risk-based framework for IFR and multi-engine training standardization in high-density airspace [25], and predictive training analytics for early identification of checkride risk, procedural noncompliance, and safety-relevant learning gaps [26]. Collectively, these studies reinforce the value of combining pilot-performance indicators, operational context, standardized training criteria, and forward-looking safety signals rather than treating human error as an isolated end-state.

Comparable work in other complex and high-risk domains also provides methodological support for context-aware prediction and structured risk fusion. Adeyekun developed an artificial-intelligence framework for predictive sediment dynamics and environmental vulnerability assessment in data-limited coastal systems [27]; Adeyekun and William examined how geological features shape biodiversity and ecosystem health [28]; and Adeyekun, Orekoya, and Abiola applied data-driven monitoring and modelling to coastal and sedimentary systems for climate resilience [29]. Within high-risk project assurance, Ashimolowo proposed a weighted computational model for risk-based quality governance [30], while Ashimolowo and Oginni linked quality governance with performance stability and strategic assurance [31]. Although these studies address different application domains, they support methodological principles relevant to aviation AI: explicit feature definition, context-sensitive modelling, transparent aggregation of multiple risk indicators, and auditable decision criteria.

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

Risk assessment in aviation-support infrastructure likewise benefits from integrating heterogeneous environmental and engineering evidence; Mu'awiya, Ayanwunmi, Sangodiji, and Adeyekun combined geotechnical and groundwater assessment for helipad construction [32]. In another real-time decision domain, Olaseinde developed protocols for rapid detection of authorized push payment fraud [33], illustrating the importance of low-latency detection when risk can escalate quickly. Olaseinde and Azeez further examined the trust implications of human reliance on hallucinating generative AI in internal-control testing [34]. These cross-domain studies are not direct evidence of pilot-state prediction; however, they support two design principles adopted in the present work: time-sensitive detection of emerging risk and retention of informed human oversight over AI-generated recommendations. Additional cross-domain examples of explicit multivariable response analysis appear in plant biotechnology and agro-ecological research. Isalar examined the role of activated charcoal and polyvinylpyrrolidone in callus quality [35] and compared auxin and cytokinin effects on callus morphology and biomass accumulation [36]. Related studies evaluated the interaction between explant developmental stage and phytohormone type in shea callogenesis [37], optimized phytohormone combinations for enhanced callus induction in medicinal-plant explants [38], linked soil nutrient dynamics with shea nut productivity and butter quality [39], assessed explant type and developmental stage in the in vitro regeneration of tropical tree species [40], and investigated sterilization strategies for controlling microbial contamination and browning in coconut inflorescence culture [41]. These plant-science studies are not aviation evidence; they are included only as methodological analogies for factorial or multivariable experimental design, transparent predictor definition, interaction analysis, response-based optimization, and traceable linkage between measured inputs and outcomes before model fusion.

The principal gap is therefore multimodal integration with strong leakage control and human-centred interpretability. Recent aviation studies have advanced integrated human-factors safety, risk-based training standardization, and predictive pilot-training analytics [24]-[26], but a transparent framework that fuses synchronized pilot-state, automation-interaction, and operational-context variables for window-level risk estimation remains necessary. The AI-HFRP framework addresses this gap by separating raw observations, engineered features, expert labels, and model outputs while retaining grouped flight-level validation and interpretable domain contributions.

III. Methodology

A. Research Framework and System Architecture

The architecture follows four stages: synchronized input acquisition, domain-level processing, human-factors risk fusion, and interpretable safety output. Human-state, automation-interaction, and operational-context measurements are converted into standardized domain indicators and then fused into a continuous probability-like risk output. Safety-event occurrence, unstable approach, go-around status, event severity, and expert labels are treated as outcomes and are excluded from pre-event predictors.

Table I Principal Data Domains used in the Study

Domain	Representative Variables
Pilot/duty	Flight hours; duty time; sleep-24 h
Human state	SP fatigue; KSS; NASA-TLX mental
Vigilance/SA	PVT RT/lapses; SAGAT accuracy
Automation	Mode transitions; awareness; confusion; surprise
Operational	Alt./speed deviation; visibility; crosswind; ATC
Outcomes	Safety event; unstable approach; severity; risk

B. Data Collection and Dataset Preparation

The supplied dataset contains 24 synchronized observations from four de-identified flight profiles. Each profile contributes one observation at take-off, climb, cruise, descent, approach, and landing.

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

The four pilot profiles are coded P017, P022, P031, and P044. Twelve observations are associated with captains and twelve with first officers. The source specification explicitly identifies the records as realistic synthetic demonstrations; therefore, the analysis is reported as a proof-of-concept evaluation rather than as airline operational evidence.

Three observations contained one missing field and 98.5% completeness; all other observations were complete. Non-applicable Mach and radio-altitude values were retained as missing rather than recoded as zero. The original dataset was not synthetically oversampled. High and Critical risk labels were combined only for binary model evaluation because the four-class distribution contained 12 Low, seven Moderate, four High, and one Critical observations.

C. Feature Engineering

Twenty pre-event numeric predictors were assigned to three conceptual domains: eight human-state indicators, seven automation-interaction indicators, and five operational-context indicators. For each cross-validation fold, continuous predictor values were standardized using the mean and standard deviation estimated only from the training flights:

$z_{j,t} = \dfrac{x_{j,t} - \mu_{j,train}}{\sigma_{j,train}}$	(1)

Higher values of sleep duration, SAGAT accuracy, and correct mode awareness represented lower risk and were therefore sign-reversed after standardization. The three domain indices were computed as signed averages, where s_k, a_k, and o_k denote the pre-specified direction of each standardized indicator:

$H_t = \dfrac{1}{8} \Sigma\, s_k\, z_{k,t}, \quad k = 1,\ldots,8$	(2)
$A_t = \dfrac{1}{7} \Sigma\, a_k\, z_{k,t}, \quad k = 1,\ldots,7$	(3)
$O_t = \dfrac{1}{5} \Sigma\, o_k\, z_{k,t}, \quad k = 1,\ldots,5$	(4)

D. Proposed AI-HFRP Risk Fusion Model

The fusion model applied fixed a priori weights rather than optimizing weights against the 24 labels. This reduces extreme overfitting and preserves interpretability. Human-state risk receives 0.45 weight, automation interaction 0.35, and operational context 0.20: The emphasis on fixed, inspectable domain contributions is also consistent with broader risk-governance work that favors explicit weighted criteria and auditable assurance structures [30], [31].

$R_t = 0.45H_t + 0.35A_t + 0.20O_t$	(5)

The fused score was transformed using a logistic function. A probability of 0.50 was used as the exploratory High/Critical warning threshold:

$P_t = \dfrac{1}{1 + e^{-R_t}}$	(6)

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

E. Benchmark Algorithms and Validation

The same 20 numeric predictors were supplied to the benchmark classifiers. Logistic Regression used standardized predictors; Random Forest used 300 trees; Support Vector Machine used a radial-basis-function kernel; Extreme Gradient Boosting used 100 trees, maximum depth 3, and learning rate 0.1; and the Artificial Neural Network used two hidden layers of 16 and 8 units. Median imputation was applied inside each training fold where required. No extensive hyperparameter search was performed because only four independent flight groups were available.

Grouped four-fold cross-validation used Flight_ID as the grouping variable so every fold withheld one complete flight profile. This prevented observations from the same flight from simultaneously appearing in training and testing data. The positive class contained the five High/Critical observations; the remaining 19 Low/Moderate observations formed the comparison class.

F. Performance Metrics and Explainability

Performance was assessed using accuracy, sensitivity, specificity, precision, F1-score, area under the receiver operating characteristic curve (AUC), and false-alarm rate. The principal expressions were:

$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	(7)
$Sensitivity = \frac{TP}{TP + FN}$	(8)
$Specificity = \frac{TN}{TN + FP}$	(9)
$F_1 = 2\frac{P \times R}{P + R}$	(10)

where P denotes precision and R denotes recall. Because the dataset contains repeated synthetic windows from only four flight profiles, Spearman rank correlations were used descriptively to identify monotonic relationships with the four-level expert label. The correlation coefficients are not treated as population-level causal estimates.

IV. Results and Discussion

A. Characteristics of the Aviation Human Factors Dataset

The 24 observations represented four complete six-phase profiles. Mean total flight experience was 5982.5 ± 2589.9 h. Mean duty duration was 6.15 ± 1.90 h, and mean sleep in the previous 24 h was 6.38 ± 0.92 h. Eight observations (33.3%) occurred in IMC. Mode confusion occurred in two observations, automation surprise in three, manual override in six, and a coded safety event in three. Two approaches were unstable and two go-arounds were recorded; no near-miss outcome was coded.

Table II Descriptive Characteristics Of Principal Variables

Variable	Mean	SD	Range
Duty time (h)	6.15	1.90	2.8-9.0
Sleep previous 24 h (h)	6.38	0.92	4.9-7.2
Samn-Perelli fatigue	3.42	1.28	2.0-6.0
KSS sleepiness	4.58	1.69	3.0-8.0
NASA-TLX mental	44.58	7.63	32.0-59.0
PVT reaction time (ms)	268.75	24.71	240.0-316.0
SAGAT accuracy (%)	86.92	6.31	73.0-94.0
Mode transitions	2.04	1.12	1.0-5.0
Cockpit alerts	5.21	3.05	1.0-11.0
Altitude deviation (ft)	121.88	84.28	40.0-420.0
Speed deviation (kt)	6.58	3.83	3.0-18.0

Volume 4 / Issue 8/ August 2026 International Journal of Modern Science and Research Technology
ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

B. Distribution of Human-Factors Risk Indicators

The expert-adjudicated distribution was imbalanced: 12 observations were Low risk, seven Moderate, four High, and one Critical. High/Critical observations therefore represented 20.8% of the dataset. This distribution is shown in Fig. 1 and is the principal reason that the four-level target was collapsed for binary classifier evaluation.

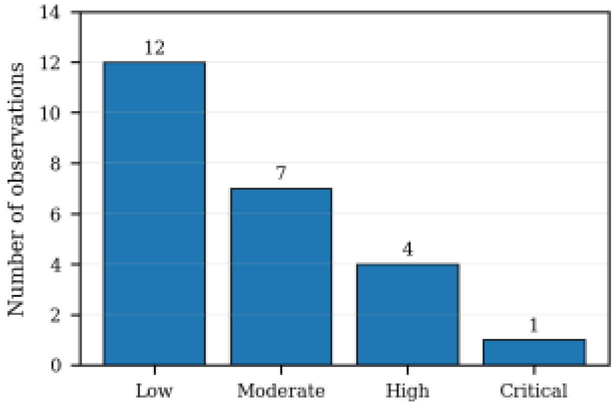


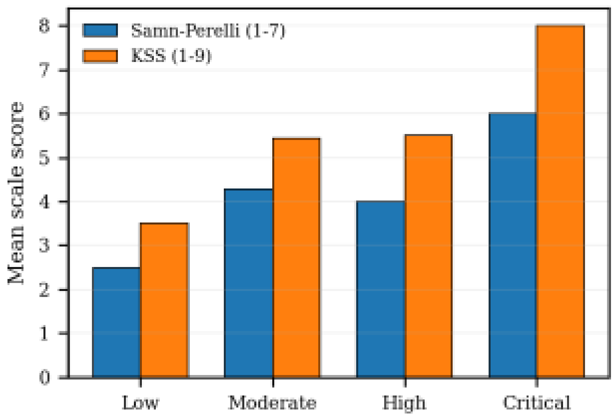
Fig. 1. Distribution of expert-adjudicated human-factors risk classes.

Table III Distribution of Adjudicated Human-Factors Risk

Risk Class	n	%
Low	12	50.0
Moderate	7	29.2
High	4	16.7
Critical	1	4.2

C. Fatigue, Workload, Vigilance, and Situation Awareness

The human-state gradient across the risk categories was clear. Mean duty duration increased from 4.77 h in Low-risk observations to 8.90 h in the Critical observation, whereas mean previous-24-hour sleep decreased from 6.95 h to 4.90 h. Samn-Perelli fatigue increased from 2.50 to 6.00 and KSS from 3.50 to 8.00. PVT reaction time increased from 252 ms in Low-risk observations to 313 ms in the Critical observation, while SAGAT accuracy decreased from 91.3% to 75.0%. Figs. 2 and 3 visualize the fatigue/sleepiness and situation-awareness patterns.



ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

Fig. 2. Mean Samn-Perelli fatigue and Karolinska Sleepiness Scale scores by risk class.

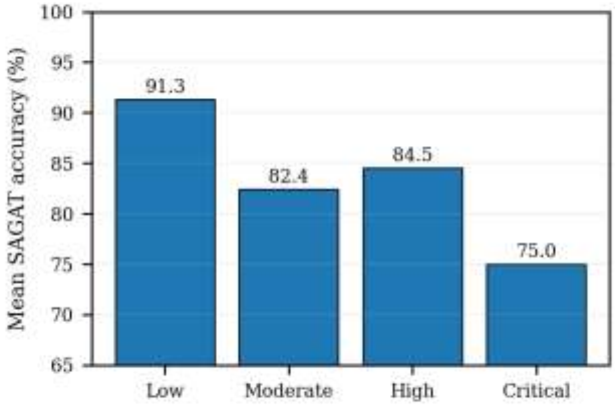


Fig. 3. Mean SAGAT situation-awareness accuracy across adjudicated risk classes.

Table IV Selected Indicators by Adjudicated Risk Class

Indicator	Low	Mod.	High	Crit.
Duty time (h)	4.8	7.2	7.8	8.9
Sleep 24 h (h)	7.0	5.8	6.0	4.9
Samn-Perelli	2.5	4.3	4.0	6.0
KSS	3.5	5.4	5.5	8.0
TLX mental	39.2	52.6	45.0	51.0
PVT RT (ms)	252.0	282.7	283.5	313.0
SAGAT (%)	91.3	82.4	84.5	75.0
Mode transitions	1.6	2.0	3.0	4.0
Alerts	3.8	7.0	5.2	10.0
Alt. deviation (ft)	90.0	111.4	201.2	260.0

D. Automation Interaction and Dominant Risk Drivers

Low-risk observations averaged 1.58 automation-mode transitions per window and retained correct mode awareness in all records. High-risk observations averaged 3.00 transitions per window. The Critical observation contained four mode transitions together with incorrect mode awareness, mode confusion, and automation surprise. This finding supports the view that automation risk is not determined by autopilot engagement alone; it is shaped by whether the flightcrew understands the active mode and can respond effectively to unexpected system behavior [9], [18].

Exploratory Spearman analysis identified duty duration, PVT reaction time, Samn-Perelli fatigue, and KSS as the strongest positive associations with increasing risk. SAGAT accuracy and previous-24-hour sleep showed strong inverse relationships. Automation surprise, lower visibility, greater crosswind, mode confusion, and incorrect mode awareness were also associated with higher risk.

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

Table V Dominant Exploratory Associations with Human-Factors Risk

Predictor	Spearman rho	Risk Direction
Duty time	0.789	Higher
PVT reaction time	0.733	Higher
Samn-Perelli fatigue	0.729	Higher
KSS sleepiness	0.704	Higher
SAGAT accuracy	-0.698	Lower
Sleep previous 24 h	-0.635	Lower
NASA-TLX mental	0.613	Higher
PVT lapses	0.613	Higher
Automation surprise	0.583	Higher
Visibility	-0.573	Lower

E. Performance of the Proposed AI-HFRP Model

The proposed AI-HFRP fusion achieved an AUC of 0.874. At the fixed 0.50 threshold, sensitivity was 0.800, specificity 0.684, precision 0.400, F1-score 0.533, and overall accuracy 0.708. The false-alarm rate was 0.316. The model therefore detected four of the five High/Critical observations but produced six false-positive warnings. This trade-off is important in aviation: high sensitivity can reduce missed hazardous states, but excessive false alerts may create nuisance-warning burden.

F. Comparative Performance

The proposed fusion produced the highest AUC among the evaluated methods. Random Forest and Extreme Gradient Boosting achieved moderate ranking performance but failed to identify a High/Critical observation at the conventional 0.50 threshold. Logistic Regression also predicted no positive observations. The Support Vector Machine identified one of the five High/Critical observations, while the Artificial Neural Network identified none. These results demonstrate why accuracy is misleading under strong class imbalance: a model can obtain approximately 79% accuracy largely by predicting the majority Low/Moderate class.

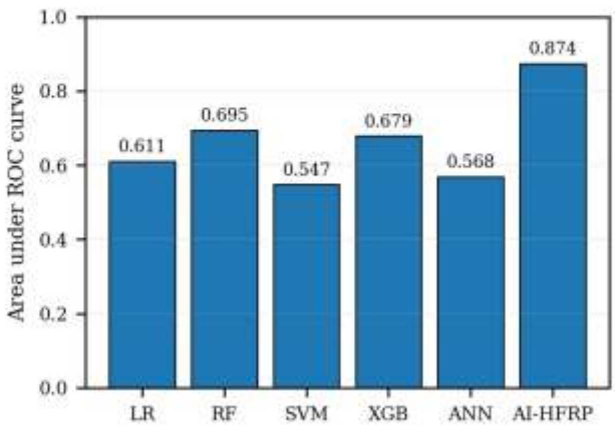


Fig. 4. Area under the receiver operating characteristic curve for benchmark classifiers and the proposed AI-HFRP fusion.

Table VI Flight-Grouped Cross-Validation Performance

Model	AUC	Sens.	Spec.	F1	FAR
Logistic Reg.	0.611	0.000	1.000	0.000	0.000
Random Forest	0.695	0.000	1.000	0.000	0.000
SVM	0.547	0.200	0.947	0.286	0.053
XGBoost	0.679	0.000	0.895	0.000	0.105
ANN	0.568	0.000	0.947	0.000	0.053
AI-HFRP	0.874	0.800	0.684	0.533	0.316

G. Confusion Matrix and Classification Performance

The pooled out-of-fold confusion matrix contained 13 true negatives, six false positives, one false negative, and four true positives. The single false negative occurred in the FLT004 descent window, for which the fusion probability was 0.496, just below the 0.50 threshold. Threshold selection and probability calibration should therefore be treated as operational design decisions rather than fixed statistical conventions.

Table VII Confusion Matrix for the Proposed Ai-Hfrp Model

	Pred. Low/Mod.	Pred. High/Crit.
Actual Low/Moderate	13	6
Actual High/Critical	1	4

H. Human-Factors Risk Across Flight Phases

Descent and approach contained the largest proportions of High/Critical observations, at 50% each. Cruise contained one High-risk observation, while take-off, climb, and landing contained no High/Critical labels. These patterns are consistent with the increased combination of vertical-path management, speed changes, configuration, ATC communication, weather assessment, runway preparation, and automation transitions during descent and approach.

Table VIII High/Critical Risk Rate By Flight Phase

Flight Phase	High/Critical n	Rate
TAKEOFF	0	0%
CLIMB	0	0%
CRUISE	1	25%
DESCENT	2	50%
APPROACH	2	50%
LANDING	0	0%

I. Real-Time Early-Warning Demonstration

The risk trajectories in Fig. 5 show how the same architecture can represent changing conditions across a flight. FLT001 remained below the 0.50 warning threshold throughout. FLT002 crossed the threshold during approach and landing despite remaining Low/Moderate by expert label, illustrating false-positive behavior. FLT003 remained above threshold throughout its profile and rose to approximately 0.96 during descent, where mode confusion, incorrect mode awareness, automation surprise, five mode transitions, five FMS actions, and a 420-ft altitude deviation were present. The following approach remained above 0.95 and was associated with manual override, a 5.6-s takeover

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

latency, ten alerts, an unstable approach, and a go-around. FLT004 increased from approximately 0.34 in cruise to 0.70 on approach as visibility decreased, crosswind and turbulence increased, and automation surprise and manual override occurred.

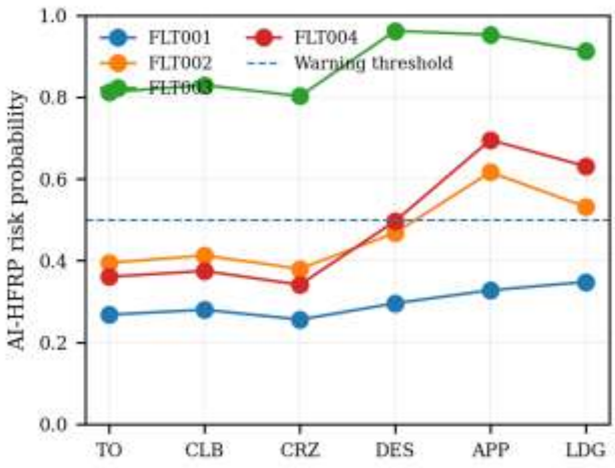


Fig. 5. Out-of-fold AI-HFRP risk-probability trajectories across the four flight profiles. The dashed line indicates the exploratory 0.50 warning threshold.

A go-around should not be interpreted automatically as a negative outcome; it can be the correct safety response to an unstable condition. The predictive objective is to identify precursor degradation before safety margins are exhausted, not to penalize appropriate recovery action.

J. Implications for Aviation Safety Management

A mature AI-HFRP system could initially support simulator training, post-flight analytics, fatigue-risk management, and SMS trend detection. It could complement FOQA by adding human-performance context to conventional flight-data exceedances. Real-time cockpit deployment would require substantially greater assurance because an alert that is poorly calibrated can itself increase workload. The safer progression is therefore from offline safety analysis and training support toward carefully validated advisory functions, with transparent explanations and retained flightcrew authority. This staged use is compatible with risk-based flight-training standardization and predictive training analytics that seek to identify competency and procedural weaknesses before they mature into safety events [25], [26].

V. Conclusion and Recommendations

A. Summary of Major Findings

The proof-of-concept analysis demonstrates that human-factors risk in highly automated operations is better represented as convergence among pilot-state, automation-interaction, and operational indicators than as a single human-error variable. Longer duty, greater fatigue and sleepiness, slower vigilance response, lower situation awareness, automation surprise, reduced visibility, crosswind, and larger flight-path deviations were associated with increasing adjudicated risk. Descent and approach contained the largest High/Critical proportions. The proposed AI-HFRP fusion achieved AUC 0.874 and sensitivity 0.800, outperforming the benchmark methods in ranking and detection of the minority High/Critical class, although the 0.316 false-alarm rate indicates a substantial calibration requirement.

B. Conclusion

AI provides a credible mechanism for moving human-factors safety analysis from predominantly retrospective event recognition toward prospective identification of deteriorating human-automation conditions. The present framework is intentionally interpretable and leakage-resistant. However, the supplied data consist of only 24 realistic synthetic windows from four flight profiles. The results therefore demonstrate technical feasibility only and do not establish accident reduction, population-level validity, or certification readiness.

C. Recommendations for Aviation Operators

Operators should evaluate AI-assisted human-performance monitoring first in simulator, training, FOQA, and SMS environments. Fatigue monitoring should combine duty and sleep context with behavioral indicators where feasible. Recurrent automation training should emphasize Flight Mode Annunciator interpretation, mode-transition recognition, FMS reprogramming under time pressure, automation reversion, and timely manual takeover. Predictive alerts should use persistence and multi-indicator logic rather than single-threshold triggers, and alert burden should be treated as a safety metric.

D. Recommendations for Manufacturers and Regulators

Manufacturers should continue improving automation transparency, mode annunciation, and visibility of system intent. Regulatory evaluation of aviation AI should require traceable data provenance, grouped and external validation, calibration analysis, false-alarm assessment, uncertainty reporting, privacy and cybersecurity controls, and explicit preservation of human authority. Models intended for safety-critical use should be evaluated across aircraft types, operators, crew compositions, environmental conditions, and abnormal scenarios. Cross-domain evidence on human reliance on unreliable generative-AI outputs further supports requiring calibrated trust, explainability, and clear human override authority for safety-relevant AI systems [34].

E. Contribution, Limitations, and Future Research

The principal contribution is an interpretable multimodal risk-fusion structure that links recognized aviation human-factors constructs to automation and operational data while maintaining separation among raw measurements, derived features, expert targets, and model outputs. The analysis also demonstrates the importance of flight-grouped validation and of evaluating sensitivity and false-alarm burden rather than accuracy alone. The major limitations are the small sample, four flight groups, synthetic proof-of-concept nature of the observations, sparse Critical class, and absence of physiological variables in the analytical records. Future studies should use substantially larger independent datasets and evaluate temporal models, physiological sensing, eye tracking, individualized baselines, probability calibration, explainable AI, digital-twin approaches, and human-AI collaborative decision support. Future work can also draw on context-aware predictive modelling in data-limited environmental systems [27]-[29], multidisciplinary aviation-support risk assessment [32], and real-time detection protocols from other high-speed risk domains [33] when designing robust multimodal data pipelines and alerting logic.

References

[1]S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds., vol. 52. Amsterdam, The Netherlands: North-Holland, 1988, pp. 139-183, doi: 10.1016/S0166-4115(08)62386-9.

[2] S. W. Samn and L. P. Perelli, *Estimating Aircrew Fatigue: A Technique with Application to Airlift Operations*, USAF School of Aerospace Medicine, Brooks AFB, TX, USA, Tech. Rep. SAM-TR-82-21, 1982.

[3] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *Int. J. Neurosci.*, vol. 52, nos. 1-2, pp. 29-37, 1990.

[4] K. Kaida et al., "Validation of the Karolinska Sleepiness Scale against performance and EEG variables," *Clin. Neurophysiol.*, vol. 117, no. 7, pp. 1574-1581, 2006, doi: 10.1016/j.clinph.2006.03.011.

[5] M. R. Endsley, "Measurement of situation awareness in dynamic systems," *Human Factors*, vol. 37, no. 1, pp. 65-84, 1995, doi: 10.1518/001872095779049499.

[6] R. M. Taylor, "Situational Awareness Rating Technique (SART): The development of a tool for aircrew systems design," in *Situational Awareness in Aerospace Operations*, AGARD-CP-478, NATO AGARD, 1990, pp. 3/1-3/17.

[7] M. Basner, D. Mollicone, and D. F. Dinges, "Validity and sensitivity of a brief Psychomotor Vigilance Test (PVT-B) to total and partial sleep deprivation," *Acta Astronaut.*, vol. 69, nos. 11-12, pp. 949-959, 2011.

[8] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cogn. Ergon.*, vol. 4, no. 1, pp. 53-71, 2000, doi: 10.1207/S15327566IJCE0401_04.
[9] Federal Aviation Administration Human Factors Team, *The Interfaces Between Flightcrews and Modern Flight Deck Systems*. Washington, DC, USA: FAA, Jun. 18, 1996.
[10] European Union Aviation Safety Agency, *Concept Paper on Artificial Intelligence, Proposed Issue 03*, 2026; and *EASA Artificial Intelligence Roadmap 2.0*, 2023.
[11] A. Alaimo, A. Esposito, C. Orlando, and A. Simoncini, "Aircraft pilots workload analysis: Heart rate variability objective measures and NASA-Task Load Index subjective evaluation," *Aerospace*, vol. 7, no. 9, Art. no. 137, 2020, doi: 10.3390/aerospace7090137.
[12] International Civil Aviation Organization, "Accident/Incident Data Reporting (ADREP) Taxonomy," ICAO Aviation Investigation and Reporting taxonomy resources, accessed 2026.
[13] Federal Aviation Administration, *Flight Operational Quality Assurance*, Advisory Circular AC 120-82, Apr. 12, 2004.
[14] NASA Aviation Safety Reporting System, "ASRS Database Online" and "About ASRS Data," NASA Ames Research Center, accessed 2026.
[15] National Transportation Safety Board, "Aviation Investigation Search/Aviation Accident Database," Washington, DC, USA, accessed 2026.
[16] D. A. Wiegmann and S. A. Shappell, *A Human Error Approach to Aviation Accident Analysis: The Human Factors Analysis and Classification System*. Aldershot, U.K.: Ashgate, 2003.
[17] J. Reason, *Human Error*. Cambridge, U.K.: Cambridge University Press, 1990.
[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A*, vol. 30, no. 3, pp. 286-297, 2000, doi: 10.1109/3468.844354.
[19] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995, doi: 10.1007/BF00994018.
[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
[22] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015, doi: 10.1038/nature14539.
[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4765-4774.
[24] O. D. T. Chukwuemeka and M. N. Mupa, "Preventing the unpredictable: An integrated human factors approach to reducing aviation accidents in the era of advanced flight technology," *IRE Journals*, vol. 9, no. 12, 2026, doi: 10.64388/IREV9I12-1719259.
[25] O. D. T. Chukwuemeka and M. N. Mupa, "Risk-based IFR and multi-engine training standardization for high-density airspace: A safety management framework for Part 141 and Part 61 flight training," *IRE Journals*, vol. 9, no. 12, 2026, doi: 10.64388/IREV9I12-1719260.
[26] O. D. T. Chukwuemeka and M. N. Mupa, "Predictive training analytics for competency-based pilot instruction: Early detection of checkride risk, procedural noncompliance and safety-relevant learning gaps," *World J. Adv. Res. Rev.*, vol. 30, no. 3, pp. 1928-1942, 2026, doi: 10.30574/wjarr.2026.30.3.1774.
[27] A. A. Adeyekun, "Artificial intelligence framework for predictive sediment dynamics and environmental vulnerability assessment in data-limited coastal systems," *J. Comput. Anal. Appl.*, vol. 33, no. 2, 2024.
[28] A. Adeyekun and J. William, "Evaluating the impact of geological features in shaping biodiversity and ecosystem health," *i TECH MAG*, vol. 7, pp. 97-101, 2025, doi: 10.26480/itechmag.07.2025.97.101.
[29] A. A. Adeyekun, O. H. Orekoya, and O. A. Abiola, "Data-driven monitoring and modelling of coastal and sedimentary systems for climate resilience," *Int. J. Data Sci. Eng.*, vol. 1, no. 1, pp. 10-42, 2023, doi: 10.34218/IJDSE_01_01_002.

ISSN:2584-2706 DOI: https://doi.org/10.5281/zenodo.22047657

[30] M. B. Ashimolowo, "A computational model for risk-based quality governance assessment in high-risk infrastructure projects: A weighted composite scoring approach," *J. Comput. Anal. Appl.*, vol. 33, no. 8, 2024.
[31] M. B. Ashimolowo and A. G. Oginni, "Quality governance and performance stability in high-risk projects: A framework for strategic assurance," *Int. J. Sci. Res. Mod. Technol.*, vol. 1, no. 2, pp. 16-40, 2022, doi: 10.38124/ijsrmt.v1i2.1282.
[32] B. A. Mu'awiya, R. Ayanwunmi, E. Sangodiji, and A. Adeyekun, "Geotechnical and groundwater assessment for helipad construction on southwestern Nigerian coastline," presented at the *GSA Connects 2025 Meeting*, San Antonio, TX, USA, 2025, doi: 10.1130/abs/2025AM-4897.
[33] A. J. Olaseinde, "The speed of theft: Developing real-time detection protocols for authorized push payment (APP) fraud in instant payment systems," *J. Comput. Anal. Appl.*, vol. 33, no. 6, 2024.
[34] A. J. Olaseinde and B. B. Azeez, "The trust paradox: Analyzing auditor reliance on hallucinating generative AI models in internal control testing," *Int. J. Artif. Intell. Eng. Transform.*, vol. 3, no. 1, pp. 38-49, 2022, doi: 10.54660/IJAIET.2022.3.1.38-49.
[35] C. E. Isalar, "Role of activated charcoal and polyvinylpyrrolidone (PVP) in enhancing callus quality in recalcitrant plants," Int. J. Sci. Res. Sci. Technol., vol. 12, no. 6, pp. 721-747, 2025, doi: 10.32628/IJSRST25126402.
[36] C. E. Isalar, "Comparative evaluation of auxins and cytokinins on callus morphology and biomass accumulation," Int. J. Sci. Res. Sci. Technol., vol. 7, no. 4, pp. 504-522, 2020, doi: 10.32628/IJSRST2293198.
[37] E. P. Aisagbonhi, C. E. Isalar, V. D. Odenore, J. U. Ogbebor, C. R. Eke, O. Asemota, and H. O. Shittu, "The interplay between explant developmental stages and phytohormone type in callogenesis of shea tree (Vitellaria paradoxa C. F. Gaertn)," Eur. Int. J. Sci. Technol., vol. 4, no. 7, pp. 50-57, 2015.
[38] C. E. Isalar, "Optimization of phytohormone combinations for enhanced callus induction in medicinal plant explants," Int. J. Sci. Res. Sci. Technol., vol. 3, no. 4, 2017, doi: 10.32628/IJSRST2316114.
[39] C. E. Isalar, "Soil nutrient dynamics and their influence on shea nut productivity and butter quality," Int. J. Sci. Res. Civ. Eng., vol. 6, no. 2, pp. 234-251, 2022, doi: 10.32628/IJSRCE226226.
[40] C. E. Isalar, "Effect of explant type and developmental stage on in vitro regeneration efficiency of tropical tree species," Int. J. Sci. Res. Sci. Technol., vol. 13, no. 4, pp. 314-326, 2026, doi: 10.32628/IJSRST2613435.
[41] A. P. Nwite, N. B. Owoh, B. O. Emoghene, C. E. Isalar, and H. O. Shittu, "Controlling microbial contamination and browning of coconut (Cocos nucifera L.) inflorescence culture," Niger. J. Biotechnol., vol. 35, no. 1, pp. 93-98, 2018, doi: 10.4314/njb.v35i1.11.