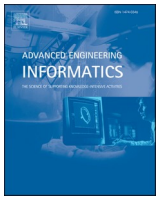
Advanced Engineering Informatics 71 (2026) 104378

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei



Enhancing aviation safety with artificial intelligence: A systematic literature review on recent advances, challenges and future perspectives☆

Cho Yin Yiu [a,b], Wen-Chin Li [b], Kam K.H. Ng [a,*], Chia-Fen Chi [c], Jens Schiefele [d]

[a] Human Factors and Ergonomics Laboratory, Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong SAR, China
[b] Safety and Accident Investigation Centre, Cranfield University, Cranfield, Bedfordshire, United Kingdom
[c] Department of Industrial Management, National Taiwan University of Science and Technology, Taipei, Taiwan
[d] Institute of Flight Systems and Automatic Control, Technical University of Darmstadt, Hessen, Germany

ARTICLE INFO

Keywords:
Deep learning
Large language models
Reliable AI
Trustworthiness
Human-AI teaming

ABSTRACT

The global air traffic is projected to grow significantly in the coming decades, leading to denser airspace and higher operational complexities. Therefore, academic and practitioners are now unleashing the potential of artificial intelligence (AI), particularly the recent advances in large language models (LLM), computer vision, and speech recognition in enhancing aviation safety through advanced cockpit design, AI assistants, human performance monitoring, and supporting air accident investigations. These applications demonstrate a significant promise in enhancing aviation safety. Nevertheless, there are still challenges in applying safe and reliable AI in supporting these safety–critical domains. Indeed, many aviation safety issues, such as accident analysis, human factors, and preventive system designs, are interconnected instead of standalone issues. This systematic literature review explores the recent advances, challenges, and future perspectives on leveraging AI to enhance aviation safety from a macro perspective. Therefore, a framework is established to review relevant studies. First, we identify the relevant literature from initial search, inspection, and screening. After that, we analyse the domains applied and the models leveraged in aviation safety enhancement on the 175 selected studies using content analysis. Then, thematic analysis is applied to reveal the challenges of applying safe and reliable AI in aviation safety. Given the challenges identified, this review recommends future work to incorporate explainable AI, develop AI certification frameworks, design based on hybrid intelligence, and adopt diversified dataset for generalisation.

1. Introduction

1.1. Background and research gap

Safety is always considered as one of the top pillars in aviation operations, given its significance in directly impacting human lives and properties both on board and on the ground. In the past decades, rigorous standard operating procedures and highly trained personnel minimised the occurrence of mechanical and human errors to keep the accident rate at a very low level. However, the surging air traffic volume increases the complexity in the airspace, which results in safety risks due to an increased workload of air traffic controllers (ATCOs) and pilots. Indeed, the annual safety report published by the International Air Transport Association [1] (IATA) reveals that the all-accident rate in 2024 of 1.13 per million flights (with seven fatal accidents) is heightened when compared with 2023 (1.09 per million flights and one fatal accident). The statistics demonstrate a rising trend of aviation accidents and an imminent necessity for advanced technological solutions to mitigate accidents at the root.

With the rapid development of artificial intelligence (AI), such as large language models (LLM), computer vision (CV), and automatic speech recognition (ASR), AI is becoming more capable of acting like human agents to analyse vast amounts of data for prediction, text classification, image recognition, and speech recognition efficiently [2]. AI has the advantage in processing a large amount of information for a faster and more accurate decision [3]. These capabilities can be

☆ This article is part of a special issue entitled: 'ADVEI_AI and LLMs in Aviation Safety' published in Advanced Engineering Informatics.

* Corresponding author.
E-mail addresses: james-cho-yin.yiu@connect.polyu.hk (C.Y. Yiu), wenchin.li@cranfield.ac.uk (W.-C. Li), kam.kh.ng@polyu.edu.hk (K.K.H. Ng), chris@mail.ntust.edu.tw (C.-F. Chi), schiefele@fsr.tu-darmstadt.de (J. Schiefele).

https://doi.org/10.1016/j.aei.2026.104378
Received 7 October 2025; Received in revised form 14 December 2025; Accepted 20 January 2026
Available online 27 January 2026

transferred to flight operations, air traffic management, and accident analysis by assisting skilled aviation professionals in their decision-making. Hence, AI has been leveraged to enhance aviation safety through active safety monitoring, unsafe flight behaviour prediction systems, and accident investigation. With the aid of generic and tailor-made AI models, a significant amount of multimodal data can be analysed more efficiently. However, it is noteworthy that AI is not designed to replace human operators. Instead, AI is designed to assist human operators in achieving safer operations [4]. Hence, the concept of human-AI teaming is proposed to utilise a joint effort from both human and AI as a co-working team to unleash the potential of a trustworthy and reliable flight operations, which should outperform the effort of either human or AI [5].

With the contemporary research efforts of leveraging AI in enhancing aviation safety, it is undeniable that AI can be a catalyst to revolutionise aviation safety from manual preventive design and accident analysis to proactive monitoring of operators, human-AI collaborative operations, and efficient accident analysis with expert systems. While AI plays a pivotal role in achieving a higher level of safety [6], Endsley [7] mentioned that there are many challenges that shall be addressed by AI-driven systems before being adopted in safety–critical domains. Existing reviews in aviation safety like [8] adopted a micro perspective to analyse how AI is applied to address a single safety issue, i.e., fatigue. However, a survey on a macro perspective can provide wider insights to address the safety issues at a systematic level. It is particularly important when many aviation safety issues, such as accident analysis, human factors, and proactive safety systems, are interconnected with each other.

1.2. Research objectives, questions, and outline

This study presents a systematic literature review on the state-of-the-art on AI applications in aviation safety from a macro perspective. In particular, we focus on the trends and the models designed/adopted by researchers, challenges faced in the technical and application levels, potential resolutions, and opportunities to strengthen AI in benefiting aviation safety in wider aspects. Specifically, we examine the literature to reveal the current technical and non-technical challenges of AI in aviation safety rather than merely focusing on the state-of-the-art models and applications. Apart from that, this study proposes a systematic literature review methodology to conduct the literature review of AI in aviation safety with both content analysis and thematic analysis. In addition, we compared the themes identified by LLM-based and human literature review to examine the potentials for collaborative literature review between human and LLM. The review lays a foundation for developing novel ideas and methodologies towards safer aviation with the aid of AI. Thus, the following four research questions (RQs) are proposed:

RQ1. *Application domains:* What are the current application domains of AI for aviation safety in research and practice?
RQ2. *Algorithms and models:* What AI algorithms and models are adopted in research and practice to enhance aviation safety?
RQ3. *Challenges and future research:* What challenges remain and how future research can promote applications of AI in aviation safety?

By answering these RQs, the status quo and recent advances of AI in aviation safety and accident investigation are revealed. The outline of the remainder of this review is as follows: **Section 2** illustrates our proposed research methodology for the systematic literature review, including the searching strategy, criteria for inclusion/exclusion. **Section 3** discusses the state-of-the-art applications of AI in aviation safety in each application domain. **Section 4** first presents the distribution and trend of algorithms and models adopted in aviation safety. Then, we summarise how different models align with the corresponding domain. The research challenges on enhancing the trustworthiness of AI for aviation safety are analysed to unveil future research directions, prospects, and opportunities towards trustworthy applications of AI in such a safety–critical domain in **Section 5**. Lastly, **Section 6** leverages LLM on literature review and cross-validates the themes identified from the literature between human and LLM's analysis. **Section 7** concludes the study with possible future work.

2. Proposed methodology of the systematic literature review

2.1. Framework

Fig. 1 illustrates the proposed systematic literature review framework for AI in aviation safety.

The proposed framework is divided into four stages: (1) literature selection, (2) content analysis, (3) thematic analysis, and (4) development of recommendations for future work. The literature selection methodology is presented in **Section 2.2**, which demonstrates the criteria that we adopted to select the literature for review. After screening, we conduct a descriptive content analysis on the manifest content to reflect the status quo of the existing research. The statistics and the general trend of publications are first presented, followed by the core domains of application and the AI algorithms/models adopted or developed for aviation safety. Then, an in-depth thematic analysis is carried out. Unlike content analysis, thematic analysis focuses on the latent content to generate themes for revealing the challenges in aviation safety. Finally, we develop several future research directions to address the captioned challenges.

2.2. Literature selection: Searching strategy and inclusion/exclusion criteria

The systematic literature review began with an online literature search. The searching criterion was formulated as: Topic = ("Artificial Intelligence" OR AI OR "Machine Learning" OR ML OR "Deep Learning" OR DL OR "Large Language Model*" OR LLM) AND Topic = (Aviation OR Flight OR Aircraft OR "Air Traffic Control") AND Topic = (Safety OR Accident). The initial literature search was conducted on Web of Science and returned 1,794 items. A preliminary inspection was first carried out, and two duplicated records were removed. Only peer-reviewed original research journal articles and conference proceedings (i.e., excluding review articles) written in English were considered. Hence, 85 items were excluded as they did not fulfil the above criterion.

A coarse-to-fine screening strategy was adopted to examine whether an item fulfils our criteria for inclusion efficiently. The primary (coarse) content screening process was subsequently conducted based on the title and abstract of each item. The title and abstract of each item were carefully reviewed to determine whether it satisfied the inclusion criteria. Articles are included if (1) they presented at least one type of AI algorithm/model, (2) the model(s) was/were applied with an objective to improve aviation safety or accident analysis. The AI algorithm/model was defined as any self-developed, self-fine-tuned, or existing LLM, deep learning (DL), and machine learning (ML) models. Articles are excluded if (1) only ideas/concepts on how to apply AI in the concerned problems, but without realisation of the proposed ideas/concepts, (2) the problem addressed was irrelevant to operational safety in aviation, such as aircraft design and maintenance. Based on the above criteria, we selected 181 items for a secondary (fine) content screening of their full text on the same set of criteria. After reading the full text of the selected items, 23 items were removed as they did not fulfil the above criteria. Finally, a forward/backward search was conducted on the items based on the citing and cited articles. Google Scholar was deployed to search for relevant articles citing the selected items. The reference list was used to perform backward searching of the articles cited by the selected items. After forward/backward searching, 17 articles were added to the selected items. The final selection comprised 175 items for review using

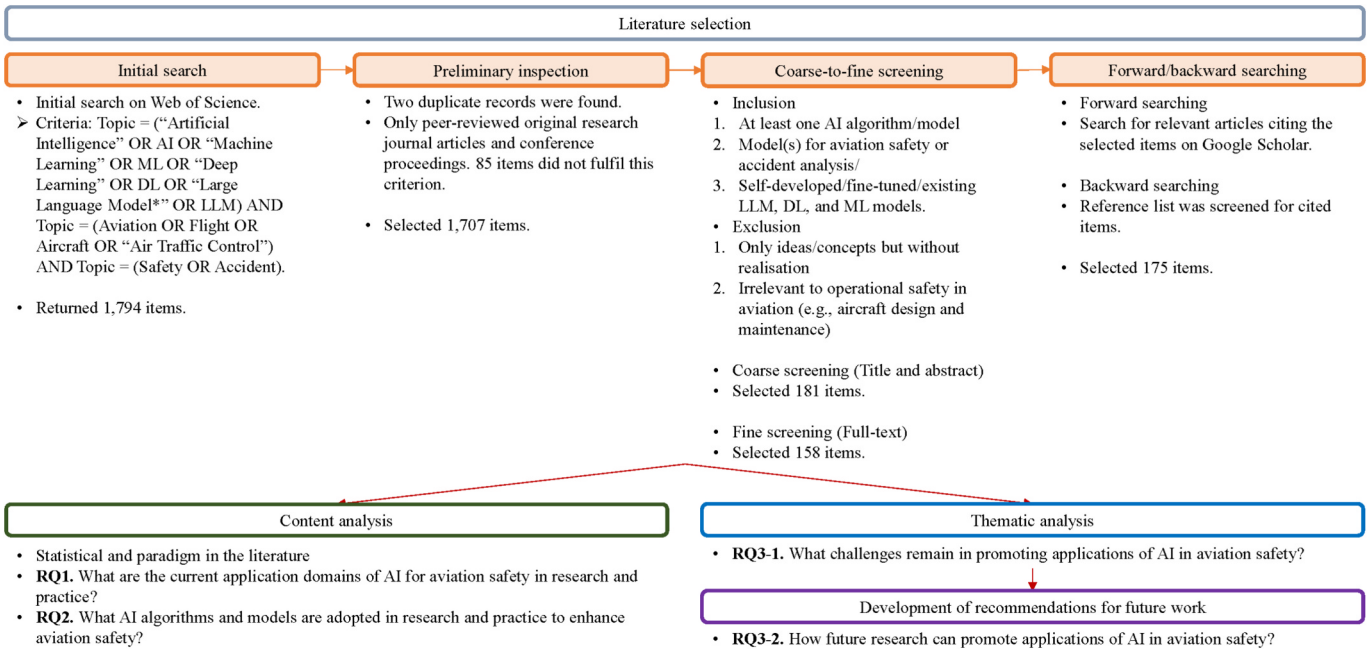


Fig. 1. Proposed systematic literature review framework.

content and thematic analyses manually.

3. What are the current application domains of AI for aviation safety in research and practice?

To identify the common application domains of AI for aviation safety, we divided this section into two main parts. First, in **Section 3.1**, we analyse the publication trend to reveal how studies in AI in aviation safety change with time. Then, we statistically summarise the distribution of publications by application domain so that readers can grasp a general picture on how AI is applied in aviation safety currently. Second, in **Section 3.2-3.4**, we introduce the related literature of each application domain to discuss the common characteristics, tasks accomplished with AI, and their corresponding findings.

3.1. Statistics and paradigm in the literature

First, we identified the trend of publication and the focus of LLM/AI-related research in aviation safety through descriptive statistics. Fig. 2 shows the number of publications by year.

From Fig. 2, the first publication related to AI in aviation safety was published in 2012. It proposed a two-stage method to identify

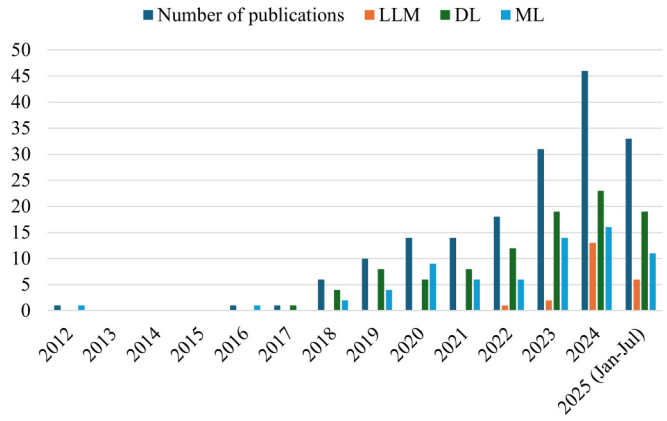


Fig. 2. Number of publications on AI in aviation safety by year.

abnormalities in flight manoeuvres using ML [9]. During the years 2013–2015, no publications were selected based on our inclusion/exclusion criteria. In 2016, Arico, Borghini, Di Flumeri, Colosimo, Bonelli, Golfetti, Pozzi, Imbert, Granger, Benhacene and Babiloni [10] designed an adaptive automation for air traffic control (ATC) based on electroencephalography (EEG), which was the first publication combining AI and neuroergonomics in aviation operations from our searching results. Then, a context-aware speech recognition model was developed for ATC applications in 2017 [11]. Since then, the number of publications using DL and ML for aviation safety has risen gradually every year. In 2022–23, more open-source LLMs with greater language abilities, such as ChatGPT 3.5 and Llama, were released and have popularised the use of LLM. Hence, the first paper leveraged LLM in aviation safety, published in 2022, developed a knowledge graph-guided and LLM-based question–answer system for aviation safety [12]. The overall number of publications reached 46 in 2024, with the surging number of LLM-related publications accounting for 28.3% of AI applications in aviation safety. Among the 175 selected items, 53.14% (93 studies) focused on flight operations. More than 30% (54 studies) were related to accident or incident analysis, identification, prediction, etc. ATC and air traffic management (ATM) accounted for 14.29% (25 studies), while a few of them jointly consider flight operations (1.71%, 3 studies).

Among the publications, accident-related publications accounted for 30.86% (54 studies) of the publications. These studies focus on accident analysis, development of a generic LLM for accident analysis, hazard identification, accident prediction, and extraction of safety concepts. More than one-third (33.71%, 59 studies) assessed human performance in flight operations and ATC/ATM. These studies designed novel AI-driven systems, including mental state classification and human performance monitoring on such as attention, operator behaviour, distraction, drowsiness, fatigue, situational awareness (SA), stress, task demand, vigilance, and workload. These studies classify mental state to ensure flight safety but are not linked to specific accidents. These studies often leveraged AI to extract and analyse the text in the accident narratives and reports. The remainder of the studies focuses on designing advanced systems for enhancing operational safety, which shared 35.43% (62 studies) of all publications. These systems are designed to enhance landing stability, identify flight risks, resolve conflicts, provide advisories and warnings to operators, construct virtual copilots to assist

operators, etc. Table 1 summarises the reviewed publications by area of application.

From Table 1, three core sub-domains were identified, i.e., sub-domains with 20 publications or above. These core specific areas included accident analysis ($n = 31$), workload identification classification and prediction ($n = 23$), and landing safety ($n = 21$).

3.2. Human performance assessment

Human factors have been a critical research domain in aviation as human intelligence remains essential in flight operations [184]. However, human subjects to many capability limitations, so researchers leverage AI to monitor human performance and provide timely corrective actions. In the following, we illustrate how AI are adopted to assess human performance with different types of data.

3.2.1. Integration of AI and neuroergonomics

Conventionally, human factors assessment heavily relied on the use of a variety of questionnaires and interviews by subject matter experts (SME) or self-reports. Particularly, various inventories were designed to assess different constructs, such as Situational Awareness Global Assessment Technique (SAGAT) and Situation Presence Assessment Method (SPAM) for SA, and NASA Task Load Index (TLX) for workload. However, these inventories could not achieve continuous monitoring, and some even require a freeze-probe that was infeasible in reality. Therefore, researchers are seeking novel tools to assess operators' performance. Among many alternatives, neuroergonomics, including EEG, electrocardiography (ECG), eye-tracking, etc., was found promising to measure human operators' SA [185], workload [186], etc. As a result, neuroergonomics was widely adopted to assess human performance with time-series data that could continuously provide insight into human mental state. Fig. 3 demonstrates a commonly adopted framework that researchers integrate neuroergonomics with AI models for human performance monitoring.

Table 1
Distribution of publications on AI for aviation safety by application domain.

Area	Count	Related publications
Accident analysis and prevention	**(54)**	
Analysis (Classification and causal factors identification)	31	[13–43]
Prevention and prediction	14	[44–57]
Hazard identification	5	[58–62]
Integration with knowledge graph	2	[12,63]
Generic LLM	1	[64]
Safety concept extraction	1	[65]
Human performance assessment	**(59)**	
Mental workload and task demand	23	[10,66–87]
Fatigue	9	[88–96]
Situational awareness	7	[97–103]
Behaviour	6	[104–109]
Attention, distraction, and inattentional deafness	4	[110–113]
Mental state (Generic)	4	[114–117]
Vigilance and Drowsiness	4	[118–121]
Human performance monitoring	1	[122]
Stress	1	[123]
Operational safety system design	**(62)**	
Landing	21	[124–144]
Anomaly detection and risk identification	10	[9,145–153]
Speech recognition	7	[11,154–159]
Separation/Conflict resolution	6	[160–165]
Flight event and safety prediction	5	[166–170]
Human-AI teaming, Single-pilot operations (SPO), and Copilot	4	[171–174]
Training	2	[175,176]
Pilot advisory and safety warning	2	[177,178]
Emergency	1	[179]
Adaptive coaching	1	[180]
Text-to-speech simulation	1	[181]
Cognitive attribute screening tool	1	[182]
Honesty behaviour prediction	1	[183]

The application of neuroergonomics in human factors was not only limited to monitoring but also to enhancing human performance through an AI-based autonomous agent that learned the physiological patterns of operators. Back in the time without AI, these complex neuroergonomic data were processed by SME for the interpretation of human mental states. However, AI provided a promising opportunity to learn complex patterns from the neuroergonomic data so that a specifically trained AI model could classify different types of mental states directly. The data-driven approach mimicked SMEs and classified human mental states based on the data provided. Therefore, since the first publication in 2016, EEG was incorporated in designing an adaptive automation for ATC [10]. With the concept of a passive brain-computer interface (pBCI), it utilised the mental workload index based on EEG data as a trigger to dynamically change the level of automation (LOA).

AI and EEG were often combined for mental workload classification and prediction [67–69,71,73,74,79,80,84–86]. Indeed, dry EEG have been proven effective to classify mental workload in real-life settings [71,80]. Other than EEG, functional near-infrared spectroscopy (fNIRS) was another key neuroimaging tool that integrates with AI to classify mental workload [76,78]. Compared to EEG, its signal quality was more stable in an operational setting, but simultaneously demanded a significantly higher initial cost, resulting in less popularity in the research. ECG was also adopted to assess mental workload based on its correlation with EEG metrics [75,77,83]. Some combined the use of EEG with eye-tracking [81] and ECG [82] to develop AI models to classify and predict cognitive workload, while eye tracking is also used independently [87]. With the development of LLM, Gao, Yue, Sun, Shan, Liu and Wu [66] further fine-tuned ChatGLM3-6B with eye-tracking gaze data to develop a WorkloadGPT for real-time workload detection. With an accuracy of 87.3%, their model demonstrated that LLM could be used for classification of cognitive workload. Other than assessing the workload of operators, SA was also a commonly assessed feature in the literature. SA refers to the understanding of the surrounding environment [187,188]. By classifying the SA of the operators, one could realise the degree to which the operators understand their current situation and provide intervention or assistance when required. In flight operations, studies mainly adopted different neuroergonomic data, including EEG [101,103], eye-tracking [98], and heart rate variability (HRV), or even a multimodal approach [100], to differentiate different levels of SA with a DL model. ATC is another key area of focus for SA classification. Similarly, multimodal neuroergonomics, including EEG and eye-tracking, is adopted [102], while Celina, Samardzic, Tukaric, Radisic and Hermann [99] adopted eye-tracking as a tool to evaluate the SA of ATCOs on their proposed conflict detection system. Fatigue was also a critical problem in aviation that might result in aviation accidents. Similarly, EEG remained the most used neuroergonomic tool for fatigue assessment with AI. Wu, Peng, Zhang, Lin and Sheng [94] adopted a deep contractive autoencoder network to recognise pilots' fatigue using a new fatigue evaluation index based on EEG band powers. Lee, Kim and Kim [92] proposed a convolutional long short-term memory (LSTM) to classify fatigue level in pilots using EEG. The multi-scale decomposition method [96] and the method of local fatigue characteristics learning from the EEG power spectrum [95] were developed to enhance the classification performance. HRV was also adopted to identify flight fatigue with LightGBM [90]. It was also combined with eye-tracking to detect fatigue in prolonged flight missions [88].

Other than the three key constructs, AI models were also applied to classify different mental states of pilots [114–117] and their levels of attention with EEG [110,111]. They were also used to identify other human factors issues that would probably result in air traffic accidents with EEG and/or eye-tracking, such as vigilance and drowsiness [118–121], as well as distraction and inattentional deafness [112,113]. Li, Li, Chen, Wang, Lu and Wen [123] detected stress of pilots using ECG, electromyography (EMG), electrodermal (EDA), respiration, and skin

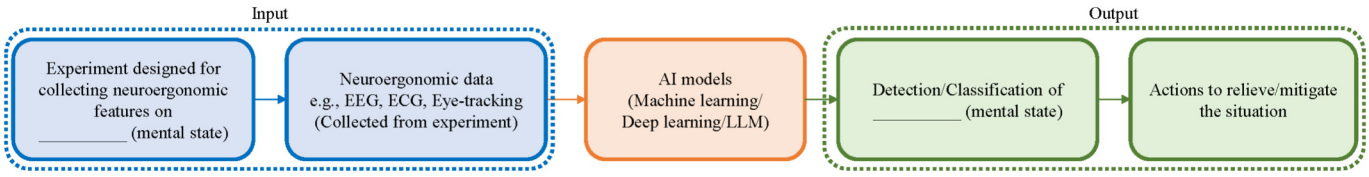


Fig. 3. A commonly adopted framework for training an AI agent with neuroergonomics on human performance monitoring.

temperature. Other than cognitive constructs, several studies also detected whether pilots had performed safe operational behaviour [104,105,107,108] and ATCOs' detection failures to warnings [109].

3.2.2. Facial recognition and automatic speech recognition

With the advancement in CV and ASR, facial recognition and speech recognition with data-driven models were made feasible with data collected from the field or experimental setting. Therefore, facial recognition models trained with pictures or videos could be used for detecting and inferring cognitive states of the human operators that were reflected on the operators' faces. For instance, with the aid of facial images collected from a simulator-based flight experiment, Shang, Si, Wang, Pan, Liu, Li, Qiu and Xu [89] integrated the face eye-mouth-face (EMF) feature model and the convolutional neural network (CNN) to detect fatigue in the flight trainees. With the superiority of CNN in image recognition, their model successfully identified 93.9% of fatigue during testing. On the other hand, Huang, Tang, Tian, Huang and Li [91] deployed an LSTM to jointly consider the facial and vocal features in detecting ATCOs' fatigue. Another study designed a GoPro-based non-contact mental workload assessment model for mental workload during landing in adverse weather [70].

Apart from facial recognition, voice-related AI also constituted a critical part in enhancing aviation safety on pilot-ATCO communication for real-time information exchange and operational guidance. ASR had the potential to visualise such speech for easier reference [156], or even to develop artificial SA for safety analysis based on the conversations [154]. In this regard, Lin, Yang, Guo and Fan [159] leveraged an LSTM to develop an ASR model for ATC with multilingual capability. They further developed a speech representation learning network to enhance recognition accuracy through capturing discriminative speech representations from raw sound waves [158]. Oualil, Klakow, Szaszak, Srinivasamurthy, Helmke and Motlicek [11] further recognised the importance of a context-aware model and developed an ASR that could dynamically integrate temporal and situational ATC context information. On the training side, Ohneiser and Ahmed [181] designed a text-to-speech (TTS) application for cost-effective, efficient, and accurate training in radio telephony communication. Nonetheless, individual differences between different operators from different backgrounds, countries/regions of origin, and cultures, could be a key factor that affects the results. Therefore, leveraging DL to learn the patterns of individual differences could yield a higher level of aviation safety. Apart from visualising speech and incorporating the operational context, Fox, Niewoehner, Rahmes, Wong and Razdan [97] leveraged LLM to process verbal communications between pilots and ATCOs for detecting anomalous situations for enhanced SA in ATC. Nevertheless, it remained essential to ensure ASR applications were safe and trustworthy for applications in the ATC setting [155,157].

3.2.3. Other non-contact approaches for human performance assessment

Apart from the neuroergonomic approach, existing research also proposed several non-contact approaches for workload prediction and fatigue detection. Pang, Hu, Lieber, Cooke and Liu [72] formulated the problem as a dynamical time-series graph classification problem and proposed a conformal-dynamical graph learning approach to predict the workload level of ATCOs. Wu and Sun [93] utilised radiotelephony communications to detect fatigue of ATCOs through a self-adaptation quantum genetic algorithm, which yielded a high accuracy of 98.5%. Similar to facial recognition and ASR mentioned in **Section 3.2.2**, these non-intrusive/non-contact approaches had the advantage of ease of implementation but also yielded a limitation in the inability to assess the underlying human performance that did not reflect ostensibly.

3.3. Learning from accidents

3.3.1. Accident classification with ML/DL

Apart from human factors, AI were also employed to analyse the cause of accidents, prevent accidents, and identify hazards and safety concepts from the lessons learnt in accident reports [64]. Accident analysis was a major application domain, as many subtasks in accident analysis could be accelerated with the aid of AI. The first study employing AI in air accident analysis was published in 2018. It combined multiple instance learning and deep recurrent neural networks (RNN) for mining multi-dimensional heterogeneous time-series data in [40]. Subsequent studies focused on post-accident analysis on extracting textual indicators for analysing accident causes [37] and aviation accident knowledge base development [30]. However, earlier models accomplished the task in a two-stage approach as they did not have the natural language processing (NLP) capabilities. The first stage transformed the reports into labelled datasets, while the second stage developed classification models with respect to adverse events. Fuller and Hook [27] analysed the impact of a safety system on accident in terms of fatal events. With the gradual development of NLP methods like semantic encoding algorithm [43], different researchers constructed ML/DL models to classify accidents and incidents [28,35], while some studies focused on a specific type of accident such as high fatality ($\geq$100) accidents [25], accidents originated from certain regions/countries [29,41], and helicopter accidents [33,36].

3.3.2. Accident analysis for causal factors identification with ML/DL

In causal factors identification, the outcome of accidents/incidents and potential aircraft damage/fatality [24] and incident causal factors [32] were identified with LSTM and attention models. Scholars also define new methods to integrate with ML/DL for accident investigation. Perboli, Gajetti, Fedorov and Lo Giudice [38] adopted semantic text similarity to automatically identify the human factors in aviation accidents. Ni, Wang, Chen and Lin [39] proposed the hybrid sampling cross-validation method to enhance the classification accuracy in causal factors identification with LightGBM. Apart from factors identification, Nanyonga, Wasswa, Turhan, Molloy and Wild [31] further developed DL models to infer aircraft damage level in a safety occurrence from text narratives. Nevertheless, the most recent studies highlighted the importance of the explainability of the accident analysis AI models, where Shapley Additive Explanations (SHAP) was used to provide global model explanations with visualisation of their feature importance [26,42].

3.3.3. Using LLM for accident analysis

To enable AI to understand the aviation accident reports or safety records for accident analysis, understanding natural language with reasoning capabilities was the prerequisite. With the advancement of human-like conversational and reasoning capabilities in LLM, many studies published in 2023 or after leveraged LLM to perform accident analysis. An earlier study first constructed an accident risk network using LLM and DL for flight training [34]. However, LLM were designed

as generic language models that were not specifically trained to understand the context and terminologies of air accident investigation. Hence, this deficiency shall be tackled with model fine-tuning with prompt engineering and other methods to gain domain-specific knowledge. Therefore, several studies leveraged LLM with aviation safety domain-specific knowledge bases to empower LLM in accident analysis tasks such as accident classification [14,16], key information and factor extraction for accelerated analysis [15,17], and text summarisation through analysing the narratives [19]. Other than NLP capability, the reasoning capability was also a critical advantage of LLM in accident analysis since accident analysis demanded a high level of reasoning to identify the root causes of accidents and human errors using Bidirectional Encoder Representations from Transformers (BERT) and data from the Aviation Safety Reporting System (ASRS) [22,23]. To achieve accurate accident analysis, the prompt design was critical [13]. Liu, Li, Ng, Han and Feng [20] combined the Human Factors Analysis and Classification System (HFACS) and Chain of Thought (CoT) prompt design to analyse accidents and concluded that LLM outperformed human experts in inferring certain types of human errors. In addition, LLM could achieve a better performance in accident analysis with knowledge graph [18] and Retrieval-Augmented Generation (RAG)-aided causal identification model [21] to extract the key details from the unstructured/semi-structured text with the National Transportation Safety Board (NTSB) reports. From the literature, we could observe that the accident analysis with AI was gradually transforming from supporting tasks like data extraction and causal factor classification using ML/DL techniques to high-level reasoning analysis using LLM, while the previous remained essential to serve as a cornerstone for subsequent reasoning analysis.

3.3.4. Accident prevention and prediction

Apart from accident analysis, it was also significant to have early accident prevention by predicting the occurrence and severity of accidents using information from safety reports [57]. Zhang and Mahadevan [53] pioneered in flight accident prediction and developed a predictive model for the severity of abnormal flight events based on risk levels. Several studies took a macro perspective to forecast the occurrence of accidents/incidents, such as the monthly accident rate [51,55], while some studies focused on a micro level to predict aircraft damage level [44,47], accident type [46], injury levels [52], flight incident rate for unsafe events [56], and fatality rate of aviation accidents [48]. Some studies merely focused on certain types of factors like human factors [50,54]. Apart from accident prediction, the prediction quality was also essential, where the problem of imbalanced classes in training data can be addressed with Bayesian optimisation [45]. Su, Sun, Peng and Guo [49] presented a two-stage approach to identify the key influencing factors of the accidents using improved Gray correlation analysis and predict accidents. This approach had the advantage to gain a better understanding of the relationship between different influencing factors to improve the prediction accuracy.

3.3.5. Hazard identification and safety concept extraction

Rather than performing large-scale accident analysis, several studies focused on a smaller scope to identify the hazards of the accidents. Based on Aircraft Communications Addressing and Reporting System (ACARS) data, Zhou, Zhuang, Zuo, Wang and Yan [58] optimised an SVM with particle swarm optimisation and LSTM to identify the in-flight hazards. Xiong, Wang, Wong and Hou [59] first leveraged BERT to process aviation maintenance reports to identify the potential hazards during operations, followed by using Bidirectional LSTM and Conditional Random Field (CRF) to extract the underlying safety hazards. Ricketts, Guo, Pelham and Barry [62] constructed a question–answer LLM for hazard identifications using an incident dataset from the Air Safety Information Management System (ASIMS). Su [60] designed a HFACS-based hazard identification system for ATC using CNN and BERT. Hou, Wang, Xiong, Zhou and Yue [61] acknowledged that uncertainty existed in hazard identification model parameters with overconfident point estimates. Therefore, they employed a Bayesian multi-scale attention CNN with a Monte Carlo dropout mechanism to estimate the internal randomness of the model to account for the uncertainty. Finally, Chandra, Jing, Bendarkar, Sawant, Elias, Kirby and Mavri [65] fine-tuned BERT with accident and incident text narratives from the NTSB and ASRS as the Aviation-BERT, which was designed to automatically extract safety concepts from the accident/incident reports to yield insights into safety.

3.3.6. Integrated knowledge graph and AI for accident analysis

A knowledge graph (KG) is a structured representation of data that describes the entities and their relationships. The use of KG provided an opportunity to integrate a large pool of scattered pieces of information and transform them into useful knowledge relationships. Particularly, LLM could couple with the constructed KG to respond to queries about the aircraft accidents. In this regard, Agarwal, Gite, Laddha, Bhattacharyya, Kar, Ekbal, Thind, Zele and Shankar [12] first constructed an Aviation KG using NTSB reports to identify the relationships between different entities in accidents. They further developed a question–answer system based on BERT and GPT-3 and use the Aviation KG to guide the LLM for response. Jing, Sawant, Bendarkar, Elias and Mavris [63] further expanded the Aviation KG with the metadata of the reports and the narratives. They adopted an Aviation-BERT [65] to classify unstructured narratives by extracting comprehensive sequences of events for advanced aviation safety analysis. Hence, the joint use of LLM and knowledge graph provided a more reliable platform for understanding the relationship of accident report entities and conducting accident analysis.

3.4. Operational safety and AI-driven systems

3.4.1. Landing safety

Landing was considered the most complicated flight phase as the time to correct or react decreased to touchdown. Unsafe landings posed significant safety risks to onboard passengers and crews and resulted in many aviation accidents. Among various unsafe landings, a hard landing was one of the commonly observed problems, which was defined as the exceedance of the touchdown limitation loading value during landing. A hard landing might result in aircraft damage and passenger injuries. Therefore, researchers focused on preventing a hard landing from happening via early detection of precursors. Most studies adopted data from the Quick Access Recorder (QAR) data [139] with LSTM [144] and multi-head self-attention transformer model [130]. Gil, Hernandez-Sabate, Enconniere, Asmayawati, Folch, Borrego-Carazo and Angel Piera [125] constructed a cockpit deployable ML and DL system to model the temporal dependencies using data from the Flight Monitoring System (FMS) for hard landing prediction. Apart from hard landing, unstable approach risk misperception and landing are also detected [128,134]. In addition, the touchdown vertical speed is also a critical factor affecting landing safety, where DASHlink data was leveraged with a probabilistic Bayesian neural network to forecast the vertical speed of aircraft at touchdown [127]. Later publications also focused on the interpretability of the model by the time steps having the highest relevance with hard landing [137], as well as the product of the attention score of the parameter and that of the time in each parameter [140].

Other studies also predicted different landing performance metrics, such as landing speed [124,138], landing pitch [126], and long landing [141], using LSTM-based models to capture the temporal dependencies for time-series prediction. Puranik, Rodriguez and Mavris [131] also applied Random Forest (RF) to estimate the true airspeed (TAS) at touchdown and extended the prediction to ground speed (GS). Several studies also focused on detecting anomalies during landing using an encoder-decoder classifier [133] and neural networks [142]. These studies provided a foundation for analysing landing safety in more complex scenarios, like the presence of wind shear [129], which

facilitated the assessment of safety risk at landing [135,136] and determined the likelihood of go-around [132]. Finally, a recent study offered a new perspective on landing safety to detect anomalies on the braking controllers that were used to decelerate the aircraft after landing [143], which was complemented with the study conducted by Dmitriev, Rhein, Beller, Broecker, Huber, Schumann and Holzapfel [179] on a safety assessment of aircraft emergency braking systems. These two studies were also an indispensable part of landing safety.

3.4.2. Anomaly detection and pilot operational risk identification

While much research focused on landing safety, identifying abnormal flight events in a timely manner was also critical in ensuring safe flight operations [149]. Smart, Brown and Denman [9] first quantified the abnormalities from flight data at a certain height to identify the flight with height at the largest difference and the corresponding flight parameters. Gao, Xu, Wang, Wu and Su [167] proposed an LSTM autoencoder to first reconstruct the inputs and adopted a constraining layer to gather the learned semantic features to make flight data outliers more distinguishable. Mangortey, Monteiro, Ackley, Gao, Puranik, Kirby, Pinon and Mavri [146] enhanced the model training by a clustering algorithm to group similar flights and identify abnormal operations that trim the large-scale dataset for abnormal flight event and risk identification, making the model computationally tractable. This approach echoed the study by Tato, Nkambou and Tato [180] that revealed models learned from several meaning flight segments could outperform those learned from the whole data. Xiong, Wang, Hou and Wong [152] identified the safety risks using a hierarchical branching CNN-Bidirectional LSTM. It assessed finer-grained risk patterns and relationships through detecting safety hazards and their associated attributes. Sun, Yang, Zhang, Jiao and Zhao [153] constructed a three-stage Probability Severity-Autoencoder-LSTM algorithm to assess the risk levels of flight events. Other than identifying high-severity flight events using QAR data [170], there were also studies that adopted publicly available data sources to identify anomalies in flight operations like automatic dependent surveillance-broadcast (ADS-B) [150], as well as some innovative approach like an energy-based approach [147].

Nevertheless, like other studies in aviation safety, the explainability of DL/ML decisions remained essential in a safety–critical domain [168]. Therefore, latent space explanations were adopted to enhance the transparency of the model, where the results was promising even if only a small portion of the data was labelled [145,148]. Li, Shang, Zheng, Wang, Liu, Liu, Li, Cao and Sun [169] further implemented the class activation mapping (CAM) method for image classification interpretation in a temporal convolutional network (TCN) to interpret flight safety predictions. Among many studies focusing on anomaly detection on abnormal flight behaviours and critical parameters [166], Xiang, Gao, Gao, Zhang and Zhang [151] adopted a different approach to identify the precursors of anomalies. They utilised multiple-instance learning to reveal the anomaly precursors so that timely corrective actions could be implemented before anomalies occurred.

3.4.3. Intelligent agents for collaborative flight operations

With the gradual development of the capability of AI, AI can now act like an intelligent agent that provides human-like conversation rather than only performing classification or regression tasks. These kinds of interaction made AI more accessible to the public and generic in responding to a variety of problems. Hence, the initiatives to achieve human-AI teaming was rising so that AI agents could act as an assistant or coworker for collaborative operations [171]. These initiatives echoed the concept of reduced crew operation (RCO) and single-pilot operations (SPO). While SPO could bring multiple benefits including reduced operation cost, while deserving greater attention on its human performance aspect – whether SPO could achieve the task requirement that was designed for two pilots, is still under discussions by different stakeholders [189,190]. Nevertheless, several studies have pioneered in designing a virtual copilot using AI and LLM to act as a copilot to assist human pilots in achieving SPO. Dong, Chen, Zhao and Wang [174] first utilised a bidirectional LSTM to model the intention tendency of pilots' behaviour in SPO. By understanding the underlying intention, safety hazards caused by inconsistent operations between pilots and cockpit automation can be avoided. Li, Feng, Yan, Lee and Ong [172] leveraged a multimodal LLM to design an automated quick procedure searching to provide real-time support during flight operations. They identified the requirements that a virtual copilot should meet, including functionality of facilitating basic collaboration, functionality of providing reliable information to pilots, ease of use, and social needs to prevent isolated feelings during long flights. The concept of virtual copilot was further integrated with neuroergonomics to design a context-aware adaptive LLM to provide different visual, auditory, and text-based cues according to the pilot's cognitive workload measured with fNIRS [173], which utilised the reasoning capabilities of LLM were utilised to provide adaptive instructions to align with the pilot's instantaneous cognitive load, focus, and procedural context. AI assistants also serves as a safety risk warning model to predict the risk and risk severity based on QAR data and provide pilots with explainable advisories [26,178].

3.4.4. Development of AI assistants for ATCOs

AI was also leveraged in many assistant systems in aviation to facilitate the work of operators. Among the AI assistants, most of them were assisting ATCOs to prevent mid-air and on-ground collision and conflicts via adherence to separation minima. Indeed, the existing conflict resolution models had deficiencies that made it hard to satisfy the real-world ATC need in terms of state and action dimensions [160]. Therefore, Janson, Ahlbrecht and Durak [161] proposed the OpenCAS based on feedforward neural networks, while Han and Huang [160] utilised reinforcement learning (RL) to let AI agents learn separation assurance policies during simulation interaction. Papadopoulos, Bastas, Vouros, Crook, Andrienko, Andrienko and Cordero [165] further enhanced graph convolutional reinforcement learning to yield high-quality conflict resolution solutions for ATCOs. Meanwhile, the temporal dynamic behaviour of flight trajectories could be captured using LSTM for conflict resolution [164]. However, these system-wise advancements on AI-based conflict detection and resolution also need to interact with human ATCOs. Stefani, Jameel, Gerdes, Hunger, Bruder, Hoemann, Christensen, Girija, Koester, Krueger and Hallerbach [162] recalled the European Union Aviation Safety Agency (EASA)'s initiative on the usage of Operational Design Domains (ODD). They defined an initial ODD as an AI-driven digital team partner of ATCOs on safety–critical tasks like conflict detection and resolution. This concept was utilised to improve the interaction between human ATCOs and the AI digital team partner [106]. To further capture the communications, intention, and behaviours, Lin, Deng, Chen, Wu, Zhang and Yang [122] constructed a multimodal ATCOs monitoring framework with ASR, intent inference, and safety monitoring using DL. These human-like AI agents could interact with human operators to formulate a team to achieve ATC tasks.

3.4.5. Human attribute screening and training

AI was also adopted in modelling other human attributes. Van Benthem and Herdman [182] proposed a cognitive health screening tool for older pilots to ensure they were fit for flying in a more efficient manner. Other than age, personal qualities such as honesty were also critical in the flight deck and ATC environment, as dishonest acts like underreporting of accidents/incidents might result in safety consequences. In this regard, Meng, Peng, Zhang, Chen, Huang, Chen and Zhao [183] adopted XGBoost to predict honest behaviours across pilots and ATCOs in different personality traits and gender, which facilitated the recruitment processes for identifying the suitable candidates.

Training was also another indispensable part of flight safety that ensured operators had gained the appropriate knowledge and qualifications that enable them to carry out their duties. Hence, Zhang, Zhang, Guo, Zhou, Wu, Yang and Lin [175] developed an automatic repetition

instruction generation using attention-based models with a real-world ATC corpus. Similarly, Lin, Wu, Guo, Zhang, Yin, Yang and Zhang [176] also created an autonomous pilot agent with sequence-to-sequence (Seq2Seq) text mapping to generate the text repetition instruction and a transformer block to enable TTS. These applications, while being the minority, were also critical to enhance flight safety from a preventive design perspective.

4. What AI algorithms and models are adopted in research and practice to enhance aviation safety?

In this section, we analyse the distribution of AI algorithms and models across different application domains. Three different types of models, including LLM, DL, and ML, are evaluated. By evaluating the model characteristics and the task nature, we can reveal the rationale of model selection. Table 2 summarise the key model, their categories, and their main application domains. Then, **Section 4.1-4.4** discuss the key models in each category to demonstrate how these commonly adopted models enhance aviation safety.

4.1. Large language models

With the gradual enhancement in computational resources in the past decade, there has been a rapid development of LLM that could process natural language and provide feedback on a variety of subject matters. In LLM, more than a billion, or even a trillion, parameters are pre-trained on a large dataset to equip computers with advanced language processing and reasoning capabilities. Unlike traditional DL/ML requires interpretation on the classification output, LLM have a unique feature of excelling at producing human-like text so that users can directly interpret the output like usual conversations. Nevertheless, most of these models are generic in nature. They have a variety of generic knowledge in multiple disciplines instead of a specific discipline. However, most safety–critical disciplines demand a high level of related expertise instead. Therefore, fine-tuning enables LLM to become capable for aviation safety and accident analysis. Fig. 4 shows the distribution of the use of different LLM as base models for fine-tuning.

Table 2
Key models of each category and their main application domains.

Models	Main application domains	Reason
Large language models		
BERT	Accident analysis	Superiority in NLP
OpenAI – GPT	Accident analysis, development of agents (virtual copilot)	Strong reasoning and conversational ability to collaborate with human operators
Meta – Llama	Accident analysis, safety prediction	Open-source and superiority in NLP
Deep learning models		
LSTM	Landing safety, flight risk prediction, human performance assessment	Capturing temporal dependency in time-series data
Neural Networks – Convolutional	Human performance assessment	Image and spatial data processing for neuroergonomic data
Encoders/Decoders Autoencoders/ Transformers/ Attention	Anomaly detection	Ability to reconstruct normal data patterns to identify anomaly
Machine learning models		
SVM	All applications as baseline	Efficient computation
Random forest	All applications	Robustness with new data
Decision Tree	Accident analysis, human performance assessment	Hierarchical for explainability
XGBoost/LightGBM/ Other gradient boosting algorithms	Accident analysis, human performance assessment	High performance in capturing complex relationships
Hybrid models		
CNN-LSTM	All applications	Spatiotemporal modelling

From Fig. 4, BERT is the most adopted model ($n = 10$) as it is considered a pioneering language model. It is built based on a transformer DL architecture trained using next sentence and masked token prediction. As an influential language model, many later LLMs are developed with reference to BERT. Hence, BERT is the most utilised model among the studies, given its longer history. The majority of these studies are dedicated to accident analysis [21,23,34,52,59,60,62]. However, the GPT family from OpenAI ($n = 7$) and the Llama family from Meta ($n = 5$) are also gaining gradual attention in aviation safety applications. In particular, the novel versions of GPT models (GPT-4 beyond) feature strong reasoning capabilities that align with the need for a trustworthy model and high-level tasks such as accident analysis and virtual copilot [20,172]. Particularly, virtual copilots require decision-making and conversational ability. The combination of these two capabilities in reasoning LLMs can address the requirement of a virtual copilot. Meanwhile, Llama models are open-source and highly customisable. With their superiority in NLP, Llama models are fine-tuned to understand the context of aviation safety with text summarisation, accident analysis, and safety prediction capabilities [15,16,18,19,64]. However, LLMs and their variant large multimodal models (LMMs) are less prevalent in human performance assessment. This phenomenon may be attributed to two key reasons: human performance assessment usually does not rely on LLM's NLP capability for text processing, and the lack of sufficient high-quality data for computationally intensive fine-tuning in large-scale models. Nevertheless, LLM has the unique feature of producing human-like text, which is promising for wider applications.

4.2. Deep learning architecture

Apart from LLM, DL models are also widely used in aviation safety. Fig. 5 shows the distribution of DL model usages. For hybrid models under multiple categories (e.g., CNN + LSTM), each of the elements is counted once in each category.

We observed from Fig. 5 that the most used DL architectures are LSTM and CNN/neural networks with convolutional layers. LSTM is a type of recurrent neural network (RNN) that has the advantage of processing time-series data. It can recognise the information in a sequence that might subsequently require and the timing at which the information was no longer required. In real-world online systems like aviation, much data is recorded in a time-series manner, so the temporal dependency of data is also critical in prediction. For instance, there can be precursors in a hard landing, such as the vertical speed of the aircraft during the landing phase. Such speed indicators are changing with time, which facilitates the prediction of landing parameters or flight risks if the temporal difference is captured by the model.

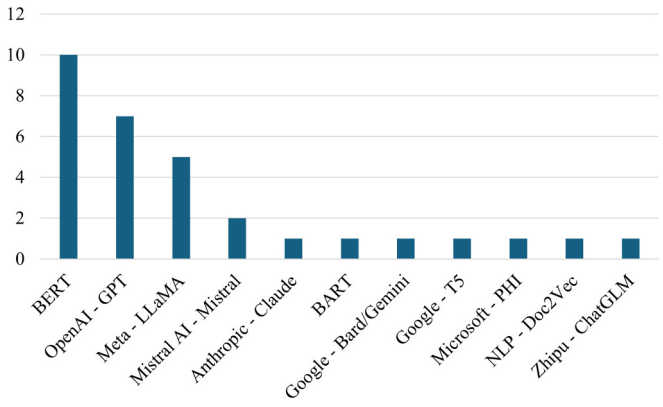


Fig. 4. Distribution of large language models adopted in different studies.

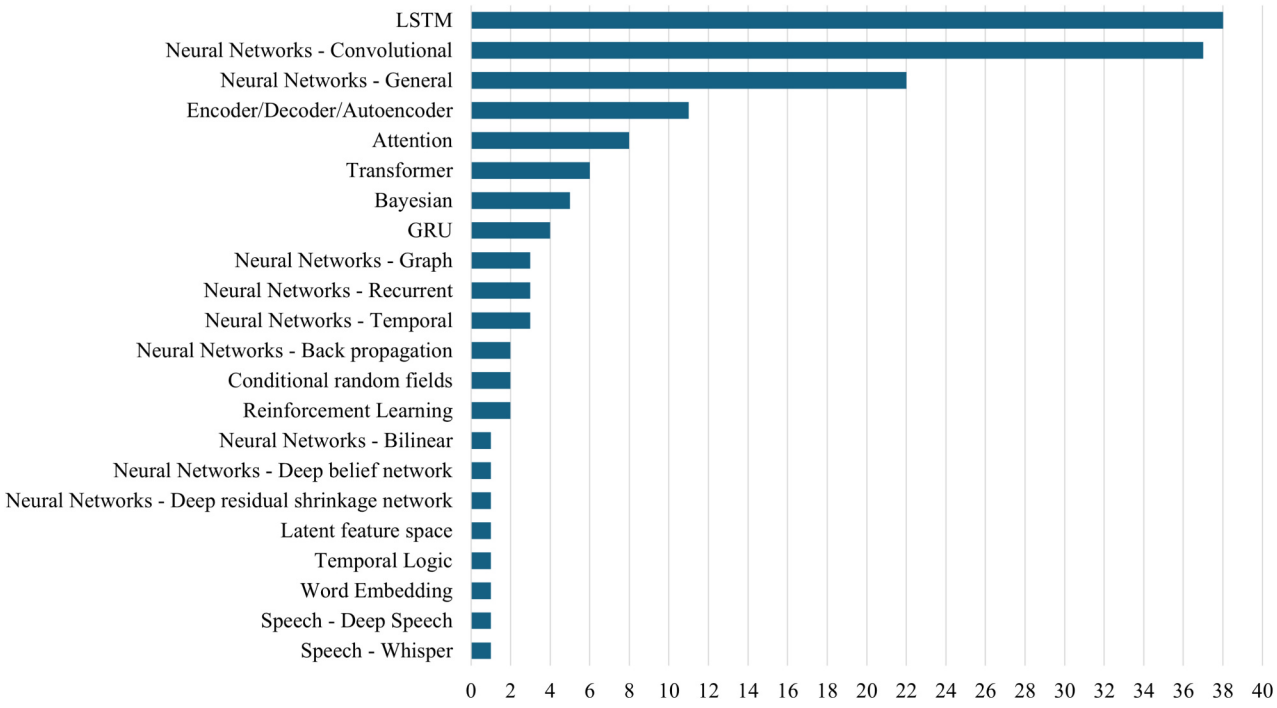


Fig. 5. Distribution of deep learning architectures adopted in different studies.

Hence, LSTM has been utilised to capture the temporal relationship of time-series flight data for predicting hard landing and landing stability [124,126,127,138,141,144]. Apart from landing safety, LSTM have also been adopted for predicting flight safety risks [152,153,167,178] and adaptive coaching [180]. Human performance monitoring is also another critical application that relies on time-series physiological data input like EEG [73,86,91,92,111,113,115,116,118,121]. Capturing the temporal dependency in these data may be essential to reveal the precursors to facilitate the prediction of safety-threatening scenarios. Conversely, as the characteristics of LSTM do not align with offline applications like accident analysis, they are not used in offline applications.

On the other hand, convolutional layers are superior in image processing as they preserve spatial relationships in the data and learned hierarchical features. Hence, CNN is widely adopted to applications requiring image processing, such as human performance assessment. Indeed, neuroergonomic data are often collected through EEG, ECG, and fNIRS, etc., to obtain physiological indicators for objective assessment of operators' state. These physiological indicators are often transformed into spatial data or images for learning. Hence, CNN are widely adopted in spatial processing of physiological data on mental states assessments of operators [68,70,72–74,78,79,83,84,89,92,110,111,113–118,121]. Indeed, many studies combine these two architectures to formulate spatiotemporal models, which will be discussed in **Section 4.4**.

Apart from spatiotemporal models and generic neural network models, the rise of advanced DL models such as encoders, decoders, autoencoders (n = 13), transformers (n = 6), and attention-based models (n = 8) is also receiving surging attention in aviation safety applications. They have superior performance in completing multimodal tasks with higher adaptability for modularised training. Among them, an autoencoder is used to reconstruct normal data patterns so that anomalies can be detected when reconstruction errors exceed a predefined threshold. In aviation, accidents and incidents are unwanted and relatively rare. Hence, developing a classification model with insufficient samples from a less prevalent class can be challenging. As a result, these advanced and recent models are thus adopted to enable anomaly detection for enhanced safety [133,143,145,148].

4.3. Machine learning models

Apart from advanced and computationally intensive DL models, conventional ML models are also the cornerstones of data-driven aviation safety models. Fig. 6 shows the distribution of the use of conventional ML models in the studies reviewed.

From Fig. 6, SVM is the most used ML model (n = 29), which many studies adopted it as a baseline model for performance comparison on risk mitigation [9,128,182], accident analysis and prediction [26,37,44,47,49,53,54,57,58], human performance prediction [67–69,71,75,77,82,85,88,90,93,102,104,105,108,109,112] given its ease of computation. It does not require heavy computational resources, which offers an efficient baseline comparison option across different application domains. Tree-based models, such as random forest (n = 24) and decision trees (n = 18), are also classical ML models that have been used for decades. While random forests are favourable for their robustness with new data, decision trees are highly explainable given their hierarchical and rule-based structure. It offers a clear visual representation of the decision-making process, which aligns with the trustworthiness requirements in aviation safety. As a result, random forest is more generic to all types of applications, while decision trees are often observed in applications requiring a higher level of explainability, like accident analysis [25,27,33,48,51] and human performance monitoring [70,71,88,90,101,109]. Hence, these supervised learning models are widely adopted in earlier ML studies in aviation safety, given their advantages in processing high-dimensional data and explainability. Nevertheless, decision trees are often prone to overfitting. Their accuracy in some complex tasks may drop when they cannot capture the complex relationships in a larger dataset. Hence, novel boosting algorithms such as XGBoost (n = 10), LightGBM (n = 5) and other gradient boosting algorithms (n = 9) are used, given their high performance and speed. While boosting algorithms have the advantage in efficiency, their performance on highly unstructured data, such as images and audio, may be limited when compared to DL models. Hence, they are mainly applied in accident analysis [33,35,36,39,41,51] and human performance assessment [75,90,107,109]. The clustering models are less used in aviation safety, but the main application is to extract or reduce features from the feature clusters. Hence, different models are applied

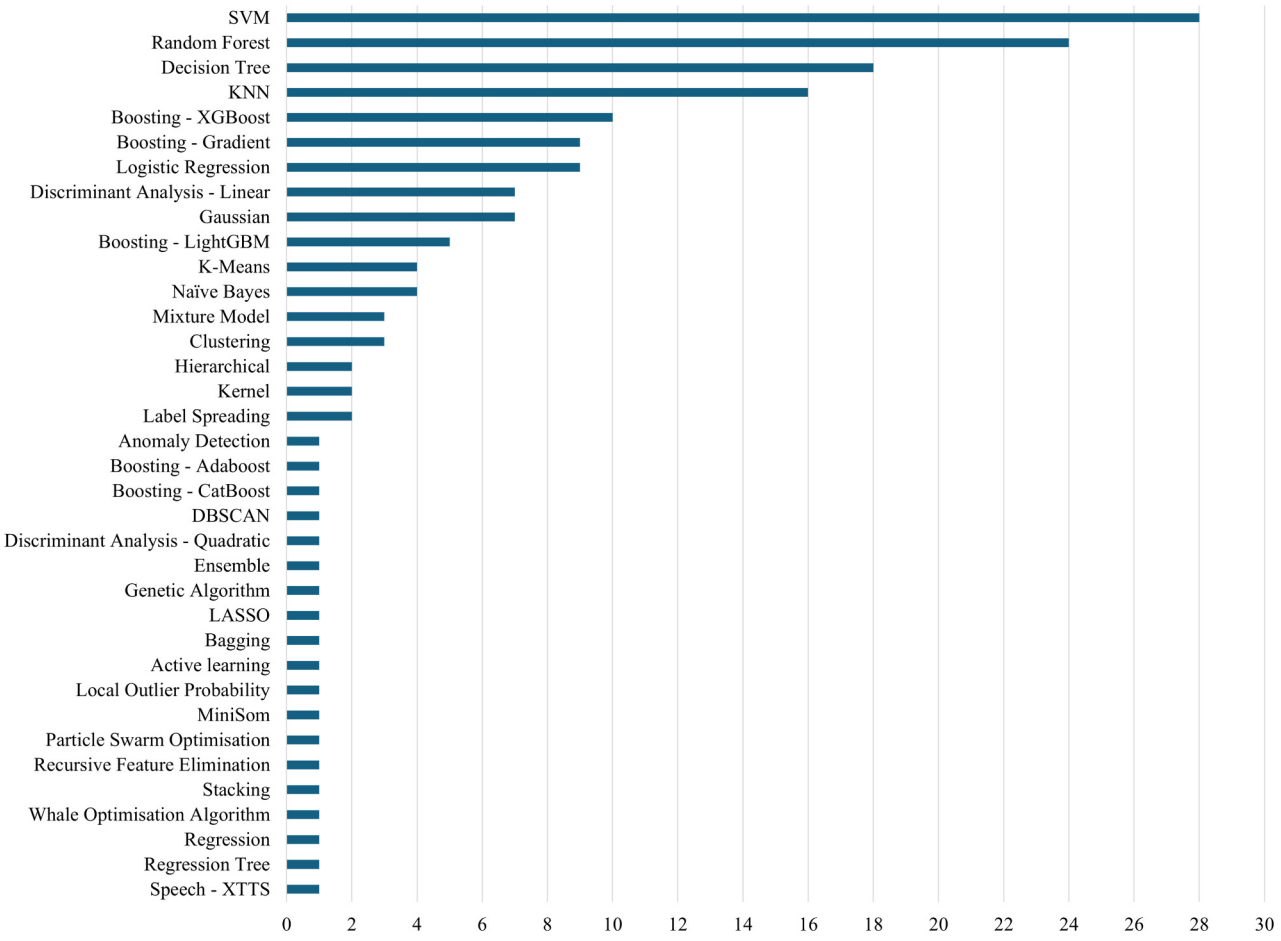


Fig. 6. Distribution of conventional machine learning models adopted in different studies.

based on the requirements so that the most appropriate model can be developed to align with the context, while these conventional models also serve as the baseline models for comparison with DL models.

4.4. Hybrid models

Indeed, instead of using a single model, multiple models and DL layers can be combined to satisfy multiple design requirements of a certain specialised context. The most common case is the combination of CNN and LSTM ($n = 13$) [28,73,92,111,113,115,116,118,121,122,141,152,158,159], which combines the spatial capability of CNN by performing convolutional operations on the data with the temporal dependency of LSTM. This combination satisfies the requirement of most AI tasks in aviation operational safety, as operational data are often changed with time and include spatial data, regardless of aircraft manoeuvres or human neuroergonomic assessments. With the gradual rise of attention-based models, convolutional layers are also combined with attention to form hybrid models for a wide range of applications like workload assessment [74,78], hazard identification [61], and hard landing prediction [140]. These hybrid models benefited from the superiority of each of the models combined, which generally yielded a higher level of accuracy.

5. What challenges remain and how future research can promote applications of AI in aviation safety?

In this section, we conducted a thematic analysis on the literature reviewed to examine the technical challenges that hinder AI and LLM in wider aviation safety applications. Based on our manual thematic analysis, we revealed three four challenges from the literature, including trustworthiness of AI models, regulatory barriers, human-AI interaction, and the dataset synchronisation, representativeness, and privacy. Pinpointing at each challenge discussed, we propose a future research paradigm to tackle such challenges and further enhance aviation safety. Furthermore, we revisited the future work proposed by each study and evaluated how their proposed future work echoes with our themes.

5.1. Trustworthiness of AI models

5.1.1. Challenges on explainability

While AI has been leveraged in multiple aspects of aviation safety, many models remain purely data-driven based on statistical relationships learned from the training data. Due to its data-driven nature, one of its main challenges is its trustworthiness [191], as a trustworthy AI should be transparent, reliable, and ethically designed for applications in safety–critical domains [192]. The lack of transparency is often described as a “black box”, indicating that the reasoning behind the decision of AI is unknown to human operators [193]. It is difficult for human operators, regulators, and investigators to trust or validate an AI model’s recommendations or actions. Additionally, bias in training data may lead to inaccurate or unfair outcomes, particularly in diverse and complex operational environments in aviation, where even fine-tuned AI may not work out the optimal solution. Wrong AI decisions may cause faulty recommendations to end-users and result in further accidents [194]. In accident analysis and prediction, it is critical to understand the intermediate process in a logical way of deriving the causes of accidents. It also aids in designing countermeasures to mitigate such accidents. On the other side, the mental state classification process with neuroergonomics is usually complex and hardly understandable by

operators. Providing explanations to operators can enhance their confidence in the outputs of AI models. In operational safety systems, the decision aids shall also be explainable so that operators can determine its trustworthiness. Summing up, "black boxes" shall be mitigated as far as possible for an informed, justified, and trustworthy decision.

Indeed, several explainable AI (XAI) approaches are developed to explain the model outputs in a post-hoc manner to enhance trustworthiness. Among XAI approaches, Shapley Additive Explanations (SHAP) is the most adopted method in aviation safety to explain the outputs of "black box" models. SHAP is a game theoretic approach that utilises Shapley values to explain the reasoning behind the prediction for a specific instance and the contributions of its predictors [195]. These Shapley values offer a fair way to distribute the overall value generated by features in all possible combinations. It aims to evaluate the impacts of each feature on the model output to explain the prediction [196]. Hence, SHAP was used to explain the model outputs on classifying mental state [114], mental workload [70,81], situational awareness [103], human out-of-the-loop [118], honest behaviour [183], and accident predictions [26,42].

While SHAP represents a significant advancement in enhancing the trustworthiness of AI models, several challenges remain that hinder their widespread application in aviation safety. A notable challenge is its computational complexity. With the number of features and samples increasing, SHAP becomes computationally expensive, especially when there are a lot of samples. It was difficult for researchers to compute Shapley values for all possible cases, limiting the validation of the model's explainability. Hence, most SHAP studies merely adopted a small sample to demonstrate the effectiveness of SHAP in explaining model output, given the long computational time. There are no unified ways to assess the explanation performance, but on a "case-by-case" basis: SHAP may explain some samples very well but simultaneously may explain some samples very poorly. Another limitation is its post-hoc nature. SHAP explains the model outputs by highlighting the features but does not explain the decision-making process of AI. Hence, the outcome of SHAP may not always align with the decision-making behind the AI model. Instead of truly interpreting how the model works internally, SHAP only approximates the reasoning behind predictions, which can still leave room for misinterpretation, especially in high-stakes domains like aviation, where absolute clarity remains essential.

5.1.2. Solution: Advancement in explainable AI

To achieve trustworthy AI in aviation safety, there is a need to advance XAI technologies to achieve global explanation performance. This is applicable to all types of applications, including accident analysis and prevention, human performance assessment, and operational safety system design. Indeed, SHAP have a high computational complexity as it relies on Shapley values to evaluate all possible feature combinations for a fair contribution assessment. Future research may focus on constructing inherently interpretable models that do not rely on post-hoc feature attribution to assess the contribution of each feature [191]. For instance, self-explainable architectures can provide both global insights into the model's logic and local explanations for specific predictions. Indeed, several research have made progress towards this direction, such as the use of graph neural networks with linear temporal logic (LTL) [119] and interlayer explanation [45]. However, the previous method generates different LTL formulas in each combination. The formula may only be able to observe a general trend in the model but not for any specific cases. The latter method does not provide implications but only explains the model structure instead of the data. Therefore, these methods can be further developed for an explainable analysis so that trustworthy explanations can be made. Furthermore, the traceability is also critical so that human operators can trace back to the original data that contributed to the prediction or action. A confidence score and reference data can be provided along with the AI's explanation to ensure every AI's decision is traceable. Lastly, another promising direction is the development of causal explainability frameworks that go beyond identifying feature contributions to uncovering causal relationships between inputs and outputs. Instead of identifying feature contributions only, the causal relationship between features/factors shall also be incorporated in the design of such AI models. Human operators shall be able to understand the causal factors to identify the logic and step-by-step thinking flow of AI, which is critical in aviation for predicting and mitigating risks.

In particular, the CoT approach to fine-tune LLM can be integrated inherently in AI model design so that the model can be self-explanatory for humans to understand their reasoning. In the following, we present an example of utilising CoT in each application domain. For accident analysis, a hierarchical approach, such as the HFACS, can be integrated with LLM prompt design. A step-by-step CoT can guide LLM to think according to the specified approach. Hence, it enhances the traceability of errors and provide a suitable explanation to each LLM action. This approach mitigates hallucinations and randomness in LLM so that human investigators can understand the logic behind LLM's response. Unlike accident analysis, there was no specific approach for human performance assessment that could be directly integrated. However, CoT can be applied by breaking down the main task (e.g., predict cognitive workload) into multiple subtasks. For instance, if EEG was adopted as the data source, LLM can be asked to compute the spectral features like theta and beta waves and comment on the results before making a conclusion. This approach helps LLM to identify the key components for identifying and predicting mental states. For operational safety, we take a hard landing as an example. Hard landing is usually characterised as the excessive vertical speed and 'g' loading. A CoT design can first help LLM to derive these parameters and determine whether abnormal values are observed. Then, LLM can conclude whether a hard landing occurred based on the values and explain its decision process accordingly. All these designs enhanced the explainability of the model to understand the decision logic. With a clearer understanding of a human-like conversation, human operators are more likely to trust these systems. Therefore, by designing inherently interpretable models, ensuring traceability in the models, and embedding causal reasoning into models, researchers could enhance the explanatory power and trustworthiness of AI systems, ensuring they align more closely with real-world aviation dynamics.

5.2. Regulatory barriers

5.2.1. Challenges on regulatory compliance

While AI holds significant promise for enhancing aviation safety, there are many challenges beyond advancing AI technologies. For instance, regulatory compliance remains the key hurdle and presents substantial challenges that must be addressed in parallel. Currently, there is a lack of standardised guidelines and evaluation frameworks for the design, implementation, and certification of AI systems in aviation safety, which leaves developers without clear pathways to compliance. The rapid evolution of AI technologies often outpaces regulatory adaptation. Indeed, AI, especially generative AI, often exhibits uncertainties and randomness in its outputs. Hallucination is one of the key concerns that AI's response presents false or misleading information as facts. It results in challenges to provide a ground truth efficiently to detect hallucinations from AI, and the way to certify whether the AI is considered acceptable in meeting stringent aviation safety standards. In complex models, it is particularly challenging to ensure transparency, accountability, and traceability in safety–critical applications. The absence of harmonised international standards further complicates the deployment of AI solutions across different jurisdictions. The accountability considerations are also a contentious issue: Who should bear the ultimate responsibility for AI's decisions? For instance, if an AI model provided an explainable yet incorrect prediction that resulted in an accident, it remains unclear how liability shall be distributed between the AI developers, end-users (operators), and regulators. Therefore, AI shall be integrated with robust regulatory oversight and traditional safety

measures to ensure comprehensive and reliable improvements in aviation safety.

5.2.2. Solution: Development of frameworks and criteria for AI certification

Currently, there is a lack of an established regulatory framework for compliance. Hence, the certification and legal framework for AI in aviation safety shall be developed. It can be divided into two key parts: the design requirement and the design certification for safe AI in aviation. The EASA [197] published the AI Roadmap 2.0, specifically designed for aviation applications. It highlighted seven requirements advocated by the EU High-Level Expert Group on AI [198]. These requirements include (1) human agency and oversight, (2) privacy and data governance, (3) diversity, non-discrimination and fairness, (4) societal and environmental well-being, (5) accountability, (6) technical robustness and safety, and (7) transparency. The EASA [197] further introduced four building blocks, starting with an AI trustworthiness analysis that characterises and assesses the safety, security, and ethics of AI. Then, AI assurance and human factors of AI encompass the development and operational explainability for developers/auditors and end users, respectively. Finally, AI safety risk mitigation acts as a follow-up approach to mitigate the residual risks caused by the 'black box'. However, these design requirements remain to be general initiatives. How can these design requirements be transformed into specific design rules/standards? How can these standards be measured? Without a well-established standard and assessment methodologies, researchers and practitioners might find it hard to achieve compliance.

Hence, future research shall consider designing a set of specific standards and evaluation methods grounded on real-world evidence. These research outcomes echo the recent public consultation by the EASA on developing future requirements for AI-based assistance and human-AI teaming [199], which provides evidence-based recommendations to shape future AI standards in aviation safety. All these efforts facilitate AI designers to design according to the standards. Nevertheless, how AI can meet stringent conventional aviation safety standards remains uncertain [172], especially whether absolute clarity is required in decisions. In addition, the consultation by the EASA only considered the operations. Further research is required to develop standards for applications in accident investigation and analysis. Particularly, the liability issues between human and AI air accident investigators when AI is employed to analyse accident causes, or even to provide safety recommendations, shall be clearly defined. These challenges underscore the need for continued research into AI certifications that are not only accurate and scalable but also interpretable, standardised, and aligned with the unique demands of aviation safety.

5.3. Human-AI interaction

5.3.1. Challenges on human-AI interface

With the outstanding task performance demonstrated in the reviewed studies, it is undoubtedly that (explainable) AI models have the capability and potential to enhance aviation safety through enhanced human performance assessment, accident analysis, support pilots to land safely, and assist ATCOs to ensure sufficient separation is made between aircraft. However, we realised that there remains a critical deficiency in the human-AI interaction from the literature. From the studies reviewed, we found that most previous research on AI/LLM in aviation safety primarily focused on what AI agents and models can be built to assist the operators. In other words, existing research focused on developing new AI 'products', i.e., new models and agents designed to address specific functional needs of pilots, ATCOs, and accident investigators in their operations, such as accident analysis, abnormal human performance detection, and hard landing prediction. While novel AI 'products' are essential, researchers often overlook an equally important aspect: how human operators adapt to these systems and how effective interaction between humans and AI agents can be achieved.

The current focus on model development neglects the complexities of human-AI collaboration, which is essential for ensuring the successful integration of AI into safety–critical aviation operations. For example, while an AI model may generate highly accurate predictions or recommendations, its functionality is limited if pilots or ATCOs cannot understand, trust, or effectively utilise the information provided by AI. In neuroergonomic studies, most studies often propose novel classification models on different mental states but lack implications on how the classification can be applied to perform necessary corrective actions to ensure flight safety. Even though classification results are available, how these results from AI can be utilised by human operators is not specified. In AI-based safety systems, rather than only providing classification/clustering results, pilots/ATCOs also need to understand how they can utilise the information/suggestion to perform the corresponding actions. For instance, if a hard landing is predicted or identified, what shall be the preventive or remedial action that can be taken to minimise the impact? In accident analysis, AI may also be subject to many limitations on the knowledge base and experience, which demand the inputs from human SMEs in certain decisions. Without a proper iteration cycle between AI's decision and human SME's comment, the recommendations suggested by AI may not be fully feasible and reliable. In this regard, rather than fully relying on AI, the effective interaction between humans and AI is thus the cornerstone. The creation of human-AI teams in analysing complex scenarios may further enhance the overall performance. Hence, it is critical to design the interface/protocol through which humans interact with AI systems.

Furthermore, some novel AI systems can be adaptive to learn from new data sources. However, human operators also need time to learn and adapt to these systems. Without proper interface designs and training, there is a risk of over-reliance on AI, where human operators accept all recommendations without critical evaluation, or under-utilisation, that human operators neglect AI outputs due to mistrust or other reasons. Nevertheless, the importance of human-AI interaction often receives insufficient attention in research.

5.3.2. Solution: Hybrid intelligence design for human-AI teaming

Upon building the human's trust in AI, future research shall also step forward to achieve human-AI teaming for aviation safety. It must be acknowledged that both humans and AI are not 'perfect': they have different expertise and deficiencies. AI is not designed to replace humans but to augment human expertise. They excel in different areas so that the strengths of both can be leveraged to enhance the overall task performance. Ultimately, hybrid intelligence can be achieved that yields a higher level of aviation safety.

In hybrid intelligence, both human and AI expertise are placed at the centre of the system design (Fig. 7). Humans and AI are designed to collaborate with each other so that a positive synergy can be achieved with the expertise from aviation safety. To achieve effective collaboration, Battiste, Lachter, Brandt, Alvarez, Strybel and Vu [200] suggested three key tenets of human-automation teaming (HAT), including transparency, bi-directional communication, and operator-directed execution. The three key tenets can be transferred to the human-AI teaming context. The first tenet of transparency is part of the

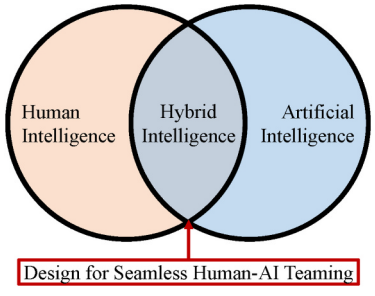


Fig. 7. Concept of hybrid intelligence.

trustworthiness issues of AI, which can be resolved by the explainable AI and AI certification strategies proposed previously. By developing accurate mental models of an AI model's functioning, operators shall be more willing to place trust in the AI systems. The second tenet, i.e., bi-directional communications, implies that human and AI shall enter a dialogue on "how to achieve their shared goals the best". To achieve so, both human operators and AI systems must first share a common understanding of the task, context, and objectives. It can be achieved by constructing a shared mental model between humans and AI so that humans can understand the reasoning, goals, and limitations of AI. Then, a proper interface can be designed to present the AI's output in a way that is intuitive and actionable for operators to facilitate bi-directional communications. Having shared mental models and human-AI interfaces, the risk of misconception and miscommunication between humans and AI can be prevented. The third tenet focuses on operator-directed execution. It indicates that while a collaborative nature between humans and AI remains, there should be only one of them to make the final decision. In certain decisions, humans may have different views from AI, leading to unresolved decisions. Hence, a key research question yet to be resolved is: How do operators perceive and respond to AI recommendations during conflicts between AI and human judgment? Future research may consider exploring the mechanisms that can resolve these kinds of conflicts to maximise the benefits of AI in hybrid intelligence so that human-AI can be teamed to yield optimal task performance.

To apply hybrid intelligence, an iterative process between human and AI shall be initiated, regardless of the application domain. For instance, in accident analysis and prediction, AI, particularly LLM, can first propose some ideas for human SME's consideration but not directly adopted. Then, human SMEs shall provide their comment based on their experience and knowledge, followed by AI revising their responses or decisions. An iterative refining process between human investigators and AIs shall form human-AI teams to accomplish the required task. The same approach can be transferred to future advanced human performance assessments with LLM. Hybrid intelligence can also be applied in operational safety system design. Taking ATC conflict resolution as an example, human controllers may find it hard to manage substantial information simultaneously. The AI-based conflict resolution can transform the presentation of information and provide necessary advisories to humans for prioritising the potential conflicts. Both intelligences are leveraged to complement each other in providing recommendations with humans remain the final decision. Eventually, this results in a human-in-the-loop AI so that humans are informed and involved for every recommendation/action, with the capability of AI leveraged to enhance the complex decision processes.

5.4. Dataset synchronisation, representativeness, and privacy

5.4.1. Challenges on dataset synchronisation

Many AI applications in aviation operational safety rely on real-time or near-real-time predictions for critical tasks, such as assessing human operators' performance (e.g., SA, cognitive workload) or predicting flight performance outcomes (e.g., hard landings, system anomalies). While these AI models often demonstrate high accuracy in offline settings, a significant challenge arises in translating these offline models into real-time operational systems. This issue stems from the inadequate consideration of data synchronisation and processing delays, which are critical for achieving the intended real-/near-time actions/feedback. Indeed, real-time AI systems in aviation require synchronised inputs from multiple data streams. For instance, EEG data are time-series in nature, but some studies are adopting the frequency domain (spectral) data as features for the classification model. To achieve so, multiple intermediate processes such as data epoching, noise removal, and Fast Fourier Transform (FFT), etc, are required but time-consuming. Nevertheless, several neuroergonomic-AI studies in aviation safety did not consider how the computationally expensive data processing work can be done dynamically in near-time to achieve the desired research outcomes. Some of these processes even heavily rely on human expertise, which prevents the system from operating autonomously in real-time. Meanwhile, the harmonisation of multiple data streams for data input requires the synchronisation between data arriving at different data rates. Similarly, in LLM studies, the scalability is also a critical challenge as LLMs are usually computationally expensive, which the difficulty of integrating these advanced models in real-time applications is a significant challenge. While these challenges can be accepted for research purposes, they hardly meet the operational requirements of real-time systems where streaming of real-time data is necessary.

5.4.2. Solution: Advanced data processing techniques and computational power

To overcome the challenges of offline prediction and data synchronisation of AI in aviation safety, future research should focus on developing innovative approaches to data processing and synchronisation. First, the data processing techniques can be enhanced by developing accurate yet lightweight algorithms for noise removal, artifact rejection, and feature extraction. Indeed, in EEG signal processing, automated artifact rejection algorithms for EEG signals were developed so that a data processing pipeline can be formed. However, the accuracy of such algorithms is questionable and often requires SME's manual review, which hinders the applications, yet is critical in supporting real-time applications of many neuroergonomic AI models in aviation safety and human factors. Hence, research may focus on constructing advanced data processing techniques with high accuracy to minimise the need for human intervention to support the implementation of application-level AI models for aviation safety. Furthermore, the initiatives on edge computing can be considered for integration with such lightweight algorithms or models for real-time applications and LLM deployment. The use of edge computing reduces the central resources required, which facilitates predictions in real-time.

Second, future research may also focus on developing online models with real-time validation. Rather than developing only classification/regression models to identify different states of operators/aircraft, future models may consider a wider perspective to design unified frameworks that consider the data flow (including data pre-processing and synchronisation) and the model's utilisation. The fusion of multiple data streams can also be done with advanced pre-processing technologies. Thus, this strategy facilitates these AI systems to be validated in real-world case studies, or at least in an online laboratory setting, which ensures the AI models developed have implications and impact on aviation safety. With the above, future AI systems can overcome the limitations of offline prediction and data synchronisation. These advancements shall enhance the operational readiness of AI systems for timely actions to enhance aviation safety.

5.4.3. Challenges on dataset representativeness and privacy

The limited diversity of datasets used to develop and validate AI models also poses a significant and recurring challenge. Due to operational limitation, many investigations rely heavily on data collected on flight simulators instead of real-world flights. In human performance assessment studies, the sample size of participants is also limited due to the challenges in recruiting qualified participants (e.g., active flying pilots), which is usually small-scale, i.e., between 10–30 participants. In addition, operators might be concerned about the data privacy in physiological monitoring. For instance, pilots might doubt if their data would be used for other purposes that affects their promotion and career. On the accident analysis side, the majority of the studies adopted NTSB reports due to the ease of access and comprehensiveness. The QAR data used are often confidential, which limits the reproducibility of the model and narrows the scope of the dataset. The lack of comprehensive and heterogeneous datasets may impede the ability to robustly assess model performance across different geographical regions and varied operational environments. In addition, the small dataset size might

cause models to be prone to overfitting.

5.4.4. Solution: Model training and testing with diversified datasets with informed consent

The current challenges of a limited diversity in datasets, particularly in physiological monitoring, arise from the mistrust of the purpose in data collection. While this challenge may not be prevalent in pure research studies, it becomes significant at the application level, i.e., when the users of the models are the employers of operators. To tackle this challenge, it is essential to establish a commonly agreed data usage policy across the industry, with informed consent obtained from operators. Operators who are being monitored shall clearly understand that their data is only used for enhancing flight safety without storing for post-evaluation purposes. In other words, only real-time applications can utilise the data for prediction, where the companies do not have access to such data afterwards. With this approach, operators can avoid data being used in an unauthorised way. It facilitates diversified data collection and promotes the models at the application level.

On the accident analysis and operational safety side, making datasets accessible by the public and more systematic by relevant civil aviation/accident investigation authorities can also promote the use of non-NTSB datasets to diversify the regional effects in the model. Indeed, most severe accident reports can be accessed online. However, these reports are usually scattered and rare thanks to the current high-level of flight safety. More diversified and systematic datasets, such as the Case Analysis and Reporting Online (CAROL) system by the NTSB, can be established by other authorities to provide more options for future studies in model training and benchmarking. The adoption of diverse datasets, if carefully selected, further enhances the class balance and strengthens the model quality. Together with the cross-validation methods like k-fold, AI models can be more robust and reliable.

5.5. Future works suggested in the studies reviewed

Apart from our thematic analysis on potential underlying challenges and future research directions, we also examined the future work suggested in the reviewed studies. Many studies suggested that there is a need to expand the dataset in terms of size and data sources ($n = 40$) and to collect real-world data for validation and implementation ($n = 23$). Indeed, many neuroergonomic studies are limited in the laboratory setting due to challenges in real-time data synchronisation. While most models achieved a high accuracy, these studies often have a small sample size, given the challenges in recruitment. For studies involving flight data, the challenge is similar as it is also difficult to sync QAR or flight data recorder (FDR) data to the model in real-time. These challenges hindered the models from being applied and creating impacts in real-world setting. Meanwhile, it aligns with the challenges in dataset representativeness and our recommendation of developing advanced techniques for real-time data synchronisation proposed in **Section 5.4**. Algorithmic enhancement is another key area of improvement mentioned in most studies ($n = 35$). While algorithmic enhancement focuses on enhancing the accuracy, most studies did not discuss how algorithms can be enhanced. Furthermore, some studies focused on certain flight phases (e.g., cruising phase) or application scenarios (e.g., limited types of accidents), which limited the scope of application of the AI model. Hence, these studies ($n = 23$) suggested expanding their scope in the future and generalising their findings to other application scenarios. In addition, numerous studies also highlighted the importance of explainability/interpretability and understanding of causal relationships and intermediate processes ($n = 20$). These studies advocated the integration of or advancement in explainable AI in future work so that operators, who are not experts in AI, can better understand the decision process behind it. It aligns with our findings on the challenge of trustworthiness on AI models and the need for advanced explainable AI technologies in **Section 5.1**. Other future works can be enhanced data quality control, better feature/parameter selection approaches, lightweight design, and validation by SMEs.

6. Comparison between human literature review and LLM-based literature review

6.1. Comparison of key insights identified between humans and LLM

Given LLM's excellence in reasoning and textual analysis, we leveraged human-AI collaboration in literature review to identify some insights to address the research questions of this review and achieve the optimal outcome. Therefore, LLM was prompted to identify the themes and insights regarding the research questions provided after our manual analysis. Models including GPT-5 and Deepseek-R1 are compared with human intelligence given their excellence in reasoning ability. In the prompt, we first set the context by telling LLM that "The attachment is a list of AI-related papers in aviation safety" and supply the publication list to LLM, including the publication titles, abstracts, and years. Then, we asked the LLM to provide insights on the research questions based on the list provided. The research questions were also included in the prompt. The prompt is included in the Appendix for reproducibility. This process was conducted after manual analysis so that the researcher would not be affected by the information provided by the LLM. Table 3 shows the mapping between themes and categories identified by humans, GPT-5, and Deepseek-R1. Similar themes are grouped into the same row for ease of reference.

6.2. Themes identified by LLM only

From Table 3, LLM revealed five themes that we did not include in our analysis, including "Explainability, trust, and assurance (GPT-5)/Explainable AI (Deepseek-R1)" in RQ2, "Validation and generalisation (GPT-5)/Generalisation and robustness (Deepseek-R1)" in RQ3, as well as "LLMs with aviation-grounded reasoning (GPT-5)/LLM specialisation and domain adaptation (Deepseek-R1)" and "Multimodal fusion at scale (GPT-5)/Multimodal and cross-modal learning (Deepseek-R1)" in RQ4.

Both LLM includes Explainable AI for RQ2 since it thinks that "Explainable models are increasingly used to improve trust and transparency", such as the use of SHAP [42] and interpretability cues [177]. However, we decided not to include "Explainable AI" as a theme in RQ2, as RQ2 focused on the algorithms and models used to learn from the training data. Explainable AI, particularly SHAP, should be viewed as a solution to accompany those algorithms and models to address the transparency challenge, which we included in RQ3 and RQ4. Hence, we decided not to follow LLM's suggestion in this case.

Both GPT-5 and Deepseek-R1 also highlighted a critical issue on the "generalisation and robustness" if data from only selected samples were trained. Therefore, a further suggestion towards future research could be incorporating diversified data during training/fine-tuning to enhance the capability of AI. Regarding domain-specific LLM, both GPT-5 and Deepseek-R1 believes that by creating a tailor-made LLM for aviation safety tasks, the domain understanding and performance can be enhanced. However, this suggestion is generic in nature and aligns with the existing LLM work in aviation safety described in **Sections 3 and 4.1**. Meanwhile, for integrating multimodal data and real-time deployment, many existing studies were already applying multiple data streams from physiological, behavioural, and operational data to generate more robust results. However, the key problem is how these multiple streams of data can be synchronised for real-time applications. Hence, both themes are the directions that shall be continued in future research instead of a novel idea for future research.

6.3. Themes identified by humans only

It is noteworthy that GPT-5 and Deepseek-R1 neglected the importance of human-AI interaction as a challenge, while GPT-5 did not identify the collaborative intelligent agents (e.g., virtual copilots) as an

Table 3
Comparison and mapping between human and LLM-based literature review.

No.	Themes identified by human (researchers)	Themes identified by GPT-5	Themes identified by Deepseek-R1
RQ1. What are the current application domains of AI for aviation safety in research and practice?			
1	(a) Accident analysis (b) Accident prediction	Incident summarisation and multi-label classification	Incident and accident analysis
2	(a) Hazard identification and safety concept extraction (b) Anomaly detection and pilot operational risk identification	(a) Causal factor extraction and coding from accident/incident narratives (b) Anomaly and outlier detection	Anomaly and risk detection
3	(a) Landing safety (b) Development of AI assistants for operators	(a) Flight data analytics for hard/long landing prediction, unstable approach detection, and braking safety (b) Real-time precursors and safety risk warning (c) Speech and communications for ATM/ATC (d) Conflict detection and ATM support	(a) Predictive safety modelling (b) Air traffic management and control
4	Integration of AI and neuroergonomics	Human performance, workload, and vigilance	Human Factors and Crew Monitoring
5	Intelligent agents for collaborative operations	/	Procedural assistance and training
6	Integrated KG and AI approach for accident analysis	Knowledge graphs and explainability	/
RQ2. What AI algorithms and models are adopted in research and practice to enhance aviation safety?			
7	Machine learning models	Classical ML	(a) Traditional ML models (b) Unsupervised and semi-supervised learning
8	Deep learning architectures	(a) Deep CNNs, sequence models, and anomaly detection (b) Graph learning and deep reinforcement learning	(a) Deep learning architectures (b) Reinforcement Learning
9	Hybrid models	Hybrid attention models and CNN-LSTM	/
10	LLM	LLM and multilingual models	LLM
11	/	Explainability, trust, and assurance	Explainable AI
RQ3. What challenges remain to promote wider applications of AI in aviation safety?			
12	(a) Trustworthiness of AI models (b) Regulatory compliance	(a) Trust, transparency, and integration (b) Explainability, accountability, and certification (c) Ethical and legal issues (d) Safety, security, reliability	(a) Interpretability and trust (b) Regulatory and certification hurdles (c) Ethical and human factor concerns
13	Dataset synchronisation, representativeness, and privacy	(a) Real-time and edge constraints (b) Data access, quality, and representativeness	(a) Real-time performance and integration (b) Data quality and availability (c) Multimodal data fusion
14	Human-AI interaction	/	/
15	/	Validation and generalisation	Generalisation and robustness
RQ4. How can future research strengthen aviation safety with enhanced AI?			
16	(a) Advancement in explainable AI (b) Development of frameworks and criteria for AI certification	(a) Advance trustworthy, certifiable AI (b) Ethics, privacy, governance, standards, and safety cases	(a) Causal and explainable AI (b) Robust and certifiable AI
17	Model training and testing with diversified datasets with informed consent	(a) Build shared, high-integrity datasets and benchmarks (b) Prospective trials and real-time deployment studies	(a) Real-time adaptive systems (b) Federated and privacy-preserving learning (c) Simulation and digital twins
18	Hybrid intelligence design for human-AI teaming	Human-centred AI and operational integration	Human-AI teaming
19	/	LLMs with aviation-grounded reasoning	LLM specialisation and domain adaptation
20	/	Multimodal fusion at scale	Multimodal and cross-modal learning

independent theme in RQ1. This phenomenon demonstrated that LLMs' reasoning ability remains limited to the information available to them without additional fine-tuning on their CoT. Indeed, the deficiency of lacking research on enhancing human-AI collaboration and teaming was not explicitly mentioned in the literature. It was identified based on a higher level of reasoning and inference during the literature review process. Hence, this outcome also suggested that LLM shall be 'educated' to gain a higher level of inference ability.

7. Conclusion

With their strong reasoning capability and efficiency, AI and LLM have been leveraged to enhance aviation safety in various ways, such as human performance monitoring, flight performance prediction, accident analysis, etc. Nevertheless, there remain many challenges for AI systems to be widely adopted in such high-stakes domains like aviation. In this study, we conducted a systematic survey to identify the core applications, models, strengths, and weaknesses in AI research for aviation safety. We discussed several research recommendations on enhancing AI and LLM for aviation safety. The contributions of this paper can be summarised as follows:

(1) Provided a holistic review of publications on AI and LLM in aviation safety: Through a review of 175 studies between 2012–2025, this paper provided an overview of how AI has been utilised to enhance aviation safety from a variety of perspectives, including operator performance monitoring, accident analysis and prediction, landing safety assessment, AI assistant development, etc. It facilitates aviation safety researchers and practitioners to understand the latest advancements and models adopted. For researchers, a thorough understanding of the state-of-the-art can facilitate novel idea to cope with the research needs in this field. For practitioners, the review provides them with an overview of the advancements of AI in aviation safety. They can consider incorporating the advancements reviewed in real-world operations and provide suggestions to researchers to refine their design.
(2) Identified the challenges and future perspectives of AI and LLM in aviation safety: While AI and LLM have been applied in aviation safety in different ways, there remain several barriers that hinder their application. Our review examines and identifies three key challenges that shall be addressed for promoting wider applications of AI and LLM in safety–critical domains. It summarised the

key problems yet to be resolved. Furthermore, we pinpointed the challenges to suggest several possible future research directions that can enhance aviation safety through advanced AI and LLM models. These perspectives cope with the identified challenges and facilitate a wider discussion and idea exchange between researchers and practitioners on enhancing aviation safety.

(3) Incorporated LLM's reasoning ability to cross-validate the themes identified: With LLM's superiority in textual analysis and reasoning, we further leverage GPT-5 and Deepseek-R1 to complement our efforts to identify the potential themes that may be overlooked by human researchers. However, it is also noteworthy that LLM also have their limitations and may overlook some themes identified by humans as well. Therefore, the proposed approach further illustrates how humans can collaborate with AI to enhance the quality of the literature review. This literature review approach can be generalised in other fields: Future literature reviews can adopt this approach to cross-validate the review outcomes while enhancing the efficiency of the literature review.

From the systematic survey, it can be concluded that AI and LLM are promising tools that can demonstrate a great impact in revolutionising aviation to data-driven approaches with intelligent human performance monitoring, AI-driven assistants on operations and accident analysis, efficient hazard identification, and ultimately achieving human-AI teaming. Nevertheless, the key challenges of AI trustworthiness and offline nature have to be tackled before AI and LLM can be safely applied in safety–critical domains like aviation. Therefore, this systematic review serves as a foundation to support future research and implementation of AI for enhancing aviation safety in a more intelligent and efficient manner without compromising the importance of the human-AI collaborative relationship.

CRediT authorship contribution statement

Cho Yin Yiu: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wen-Chin Li:** Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Kam K.H. Ng:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Chia-Fen Chi:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jens Schiefele:** Writing – review & editing, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The research is supported by Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong SAR. Our gratitude is also extended to the Research Committee of the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University for support of the project (RLPA). Cho Yin Yiu is a recipient of the Hong Kong PhD Fellowship (Reference number: PF21-62058).

Appendix. Prompt used to derive themes in LLM-based literature review

The attachment is a list of AI-related papers in aviation safety. Based on the information provided, can you please tell me about some insights on:

RQ1. What are the current application domains of AI for aviation safety in research and practice?

RQ2. What AI algorithms and models are adopted in research and practice to enhance aviation safety?

RQ3-1. What challenges remain to promote wider applications of AI in aviation safety?

RQ3-2. How can future research strengthen aviation safety with enhanced AI?

Data availability

Data will be made available on request.

References

[1] International Air Transport Association, IATA Annual Safety Report 2024, 2025.
[2] B. Wu, R. Xiao, Evolutionary attraction–repulsion algorithm embedded with LLM for UAV task allocation, Adv. Eng. Inform. 66 (2025) 103428.
[3] H. Ali, D.-T. Pham, S. Alam, M. Schultz, M.Z. Li, Y. Wang, E. Itoh, V.N. Duong, Human-AI hybrids in safety-critical systems: concept, definition and perspectives from air traffic management, Adv. Eng. Inform. 65 (2025) 103256.
[4] T.B. Sheridan, R. Parasuraman, Human-automation interaction, Rev. Hum. Factors Ergon. 1 (2005) 89–129.
[5] S. Rothfuß, M. Wörner, J. Inga, A. Kiesel, S. Hohmann, Human–machine cooperative decision making outperforms individualism and autonomy, IEEE Trans. Hum.-Mach. Syst. 53 (2023) 761–770.
[6] D. Li, A. Yao, K. Feng, H. Zhou, R. Wang, M. Cheng, H. Li, D. Wang, S. Ding, Next frontiers of aviation safety: system-of-systems safety, Engineering 52 (2025) 262–277.
[7] M.R. Endsley, Ironies of artificial intelligence, Ergonomics (2023) 1–13.
[8] S. Xin, K.Y.H. Lim, M.-H. Hsieh, C.-H. Chen, L. Dong, Managing the fatigue frontier: AI applications in air traffic control operations, Adv. Eng. Inform. 68 (2025) 103660.
[9] E. Smart, D. Brown, J. Denman, A two-phase method of detecting abnormalities in aircraft flight data and ranking their impact on individual flights, IEEE Trans. Intell. Transp. Syst. 13 (2012) 1253–1265.
[10] P. Arico, G. Borghini, G. Di Flumeri, A. Colosimo, S. Bonelli, A. Golfetti, S. Pozzi, J.-P. Imbert, G. Granger, R. Benhacene, F. Babiloni, Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment, Front. Hum. Neurosci. 10 (2016) 13.
[11] Y. Oualil, D. Klakow, G. Szaszak, A. Srinivasamurthy, H. Helmke, P. Motlicek, A context-aware speech recognition and understanding system for air traffic control domain, IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, Okinawa, Japan, 2017, pp. 404–408.
[12] A. Agarwal, R. Gite, S. Laddha, P. Bhattacharyya, S. Kar, A. Ekbal, P. Thind, R. Zele, R. Shankar, Knowledge Graph - Deep Learning: A Case Study in Question Answering in Aviation Safety Domain, 13th International Conference on Language Resources and Evaluation (LREC), Marseille, France, European Language Resources Assoc-Elra, 2022, pp. 6260–6270.
[13] P. Bert, La pression barométrique: recherches de physiologie expérimentale, G, Masson, 1878.
[14] Q. Li, K.K.H. Ng, S.C.M. Yu, C.Y. Yiu, F. Li, F.T.S. Chan, Using EEG and eye-tracking as indicators to investigate situation awareness variation during flight monitoring in air traffic control system, J. Navig. 77 (2025) 485–506.
[15] C.Y. Yiu, K.K.H. Ng, Q. Li, X. Yuan, Gaze behaviours, situation awareness and cognitive workload of air traffic controllers in radar screen monitoring tasks with varying task complexity, Int. J. Occup. Saf. Ergon. 31 (2025) 504–515.
[16] C. Zhang, J. Yuan, Y. Jiao, H. Liu, L. Fu, C. Jiang, C. Wen, Variation of pilots' mental workload under emergency flight conditions induced by different equipment failures: a flight simulator study, Transp. Res Record 2678 (2024) 365–377.
[17] W. Zhu, Y. Xie, Y. Wang, C. Zhang, J. Yuan, H. Chen, X. Zuo, C. Jiang, T. Wang, Classification of carrier-based aircraft pilot mental workloads based on feature-level fusion and decision-level fusion of PPG and EEG signals, Aeronaut. J. 20 (2025).
[18] H.T. Gorji, N. Wilson, J. VanBree, B. Hoffmann, T. Petros, K. Tavakolian, Using machine learning methods and EEG to discriminate aircraft pilot cognitive workload during flight, Sci. Rep. 13 (2023) 13.
[19] L. Salvan, T.S. Paul, A. Marois, Dry EEG-based Mental Workload Prediction for Aviation, IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), IEEE, Barcelona, Spain, 2023.

[20] D.-H. Lee, S.-J. Kim, S.-H. Kim, Decoding EEG-based Workload Levels Using Spatio-temporal Features Under Flight Environment, 12th International Winter Conference on Brain-Computer Interface (BCI), IEEE, 2024.
[21] Y. Zhou, J. Jiang, L. Wang, S. Liang, H. Liu, Enhanced cognitive load detection in air traffic control operators using EEG and a hybrid deep learning approach, IEEE Access 13 (2025) 12127–12137.
[22] Y. Wang, M. Han, Y. Peng, R. Zhao, D. Fan, X. Meng, H. Xu, H. Niu, J. Cheng, T. Liu, LGNet: Learning local-global EEG representations for cognitive workload classification in simulated flights, Biomed. Signal Process. Control 92 (2024) 13.
[23] J.A. Blanco, M.K. Johnson, K.J. Jaquess, H. Oh, L.-C. Lo, R.J. Gentili, B. D. Hatfield, Quantifying cognitive workload in simulated flight using passive, dry EEG measurements, IEEE Trans. Cogn. Dev. Syst. 10 (2018) 373–383.
[24] A. Hernandez-Sabate, J. Yauri, P. Folch, M.A. Piera, D. Gil, Recognition of the mental workloads of pilots in the cockpit using EEG signals, Appl. Sci. 12 (2022) 14.
[25] C. Liu, C. Zhang, L. Sun, K. Liu, H. Liu, W. Zhu, C. Jiang, Detection of pilot's mental workload using a wireless EEG headset in airfield traffic pattern tasks, Entropy 25 (2023) 24.
[26] G. Jiang, H. Chen, C. Wang, P. Xue, Mental workload artificial intelligence assessment of pilots' eeg based on multi-dimensional data fusion and LSTM with attention mechanism model, Int. J. Pattern Recognit Artif Intell. 36 (2022) 19.
[27] C. Zhang, S. Luo, S. Cao, Y. Zhang, H. Chen, C. Jiang, Y. Zhou, Evaluating pilot mental workload using fNIRS-based functional connectivity features with a deep residual shrinkage network under emergency flight scenarios, Int. J. Hum.-Comput. Interact. (2024) 16.
[28] C. Zhang, C. Jiang, Y. Xie, S. Cao, J. Yuan, C. Liu, W. Cao, Y. Li, Assessing pilot workload during takeoff and climb under different weather conditions: a fNIRS-based modeling using deep learning algorithms, IEEE Trans. Aerosp. Electron. Syst. 61 (2025) 1705–1724.
[29] Z. Jiang, K. Zhang, K. Wu, J. Xu, X. Li, Y. Sun, X. Ge, M. Mao, Mental workload recognition using ECG and machine learning in simulated flight tasks, 6th IEEE Advanced Information Technology, Electronic and Automation Control Conference (IEEE IAEAC), IEEE, Beijing, China, 2022, pp. 1560–1565.
[30] Y. Wang, C. Zhang, C. Liu, K. Liu, F. Xu, J. Yuan, C. Jiang, C. Liu, W. Cao, Analysis on pulse rate variability for pilot workload assessment based on wearable sensor, Hum. Factors Ergonom. Manuf. Serv. Ind. 34 (2024) 635–648.
[31] P. Xi, A. Law, R. Goubran, C. Shu, Pilot Workload Prediction from ECG Using Deep Convolutional Neural Networks, IEEE International Symposium on Medical Measurements and Applications (IEEE MeMeA), IEEE, Istanbul, Turkey, 2019.
[32] X. Yu, C.-H. Chen, H. Yang, Cognitive workload quantification for air traffic controllers: an ensemble semi-supervised learning approach, Adv. Eng. Inform. 64 (2025) 10.
[33] S.G. Hajra, P. Xi, A. Law, A comparison of ECG and EEG metrics for in-flight monitoring of helicopter pilot workload, IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE (2020) 4012–4019.
[34] B. Liu, S.W. Lye, K.X. Yeo, C.-H. Chen, A human-centric model for task demand assessment based on unsupervised learning-assisted eye movement measure, Adv. Eng. Inform. 65 (2025) 18.
[35] Y. Gao, L. Yue, J. Sun, X. Shan, Y. Liu, X. Wu, WorkloadGPT: a large language model approach to real-time detection of pilot workload, Appl. Sci. 14 (2024) 28.
[36] N. Liang, J. Yang, D. Yu, K.O. Prakah-Asante, R. Curry, M. Blommer, R. Swaminathan, B.J. Pitts, Using eye-tracking to investigate the effects of pre-takeover visual engagement on situation awareness during automated driving, Accid. Anal. Prev. 157 (2021) 106143.
[37] M.R. Endsley, Toward a theory of situation awareness in dynamic systems, Hum. Factors 37 (1995) 32–64.
[38] C. Feng, S. Liu, X. Wanyan, Y. Dang, Z. Wang, C. Qian, β-wave-based exploration of sensitive EEG features and classification of situation awareness, Aeronaut. J. 128 (2024) 2561–2576.
[39] C.Y. Yiu, K.K.H. Ng, X. Li, X. Zhang, Q. Li, H.S. Lam, M.H. Chong, Towards safe and collaborative aerodrome operations: assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks, Adv. Eng. Inform. 53 (2022) 19.
[40] G.J.W. Xu, S. Pan, P.Z.H. Sun, K. Guo, S.H. Park, F. Yan, M. Gao, X. Wanyan, H. Cheng, E.Q. Wu, Human-factors-in-aviation-loop: multimodal deep learning for pilot situation awareness analysis using gaze position and flight control data, IEEE Trans. Intell. Transp. Syst. 26 (2025) 8065–8077.
[41] C. Qian, S. Liu, X. Wanyan, C. Feng, Z. Li, W. Sun, Y. Wang, Situation awareness discrimination based on physiological features for high-stress flight tasks, Aerospace 11 (2024) 21.
[42] Q. Li, K.K.H. Ng, S.C.M. Yu, C.Y. Yiu, M. Lyu, Recognising situation awareness associated with different workloads using EEG and eye-tracking features in air traffic control tasks, Knowledge-Based Syst. 260 (2023) 16.
[43] V. Celina, K. Samardzic, I. Tukaric, T. Radisic, R.H. Hermann, Gaze Analysis of Air Traffic Contoller Using AI-Based Conflict Detection, 43rd AIAA DATC/IEEE Digital Avionics Systems Conference, IEEE, San Diego, CA, 2024.
[44] E.Q. Wu, X.Y. Peng, C.Z. Zhang, J.X. Lin, R.S.F. Sheng, Pilots' fatigue status recognition using deep contractive autoencoder network, IEEE Trans. Instrum. Meas. 68 (2019) 3907–3919.
[45] D.-H. Lee, S.-J. Kim, S.-H. Kim, Decoding Fatigue Levels of Pilots Using EEG Signals with Hybrid Deep Neural Networks, 13th International Conference on Brain-Computer Interface (BCI), IEEE, 2025.
[46] Y. Chu, Q. Wu, Recognition of Fatigue Status of Pilots Based on Deep Contractive Sparse Auto-encoding Network, 37th Chinese Control Conference (CCC), IEEE, Wuhan, China, 2018, pp. 9220–9225.
[47] E.Q. Wu, P.-Y. Deng, X.-Y. Qiu, Z. Tang, W.-M. Zhang, L.-M. Zhu, H. Ren, G.-R. Zhou, R.S.F. Sheng, Detecting fatigue status of pilots based on deep learning network using EEG signals, IEEE Trans. Cogn. Dev. Syst. 13 (2021) 575–585.
[48] D. Guo, C. Wang, Y. Qin, L. Shang, A. Gao, B. Tan, Y. Zhou, G. Wang, Assessment of flight fatigue using heart rate variability and machine learning approaches, Front. Neurosci. 19 (2025) 9.
[49] H. Qin, X. Zhou, X. Ou, Y. Liu, C. Xue, Detection of mental fatigue state using heart rate variability and eye metrics during simulated flight, Hum. Factors Ergonom. Manuf. Serv. Ind. 31 (2021) 637–651.
[50] I. Alreshidi, D. Bisandu, I. Moulitsas, Illuminating the neural landscape of pilot mental states: a convolutional neural network approach with shapley additive explanations interpretability, Sensors 23 (2023) 20.
[51] D.-H. Lee, J.-H. Jeong, B.-W. Yu, T.-E. Kam, S.-W. Lee, Autonomous system for EEG-based multiple abnormal mental states classification using hybrid deep neural networks under flight environment, IEEE Trans. Syst. Man Cybern. -Syst. 53 (2023) 6426–6437.
[52] S.-Y. Han, N.-S. Kwak, T. Oh, S.-W. Lee, Classification of pilots' mental states using a multimodal deep learning network, Biocybern, Biomed. Eng. 40 (2020) 324–336.
[53] D.-H. Lee, J.-H. Jeong, K. Kim, B.-W. Yu, S.-W. Lee, Continuous EEG decoding of pilots' mental states using multiple feature block-based convolutional neural network, IEEE Access 8 (2020) 121929–121941.
[54] T. Xu, Y. Sun, Z. Zeng, Y. Guo, An EEG-Based Pilots' Attention Detection Method, 2024 International Conference on Control and Robotics, IEEE, Yokohama, Japan, 2024, pp. 366-370.
[55] Q.A. Nguyen, N.A. Dao, L. Nguyen, Enhanced pilot attention monitoring: a time-frequency EEG analysis using CNN-LSTM networks for aviation safety, Information 16 (2025) 31.
[56] C.Y. Yiu, K.K.H. Ng, Q. Li, X. Yuan, Keeping pilots in the loop: an explainable spatiotemporal EEG-driven deep learning framework for adaptive automation in cruising flight phase, IEEE Trans. Intell. Transp. Syst. 26 (2025) 9838–9851.
[57] M. Lyu, F. Li, C.-H. Lee, C.-H. Chen, VALIO: Visual attention-based linear temporal logic method for explainable out-of-the-loop identification, Knowledge-Based Syst. 299 (2024) 14.
[58] A. Ghaderi, F. Saghafi, Enhancing pilot vigilance assessment: the role of flight data and continuous performance test in detecting random attention loss in short IFR flights, J. Air Transp. Manag. 120 (2024) 11.
[59] J.-H. Jeong, B.-W. Yu, D.-H. Lee, S.-W. Lee, Classification of drowsiness levels based on a deep spatio-temporal convolutional bidirectional LSTM network using electroencephalography signals, Brain Sci. 9 (2019) 18.
[60] E. Masse, O. Bartheye, L. Fabre, Classification of electrophysiological signatures with explainable artificial intelligence: the case of alarm detection in flight simulator, Front. Neuroinf. 16 (2022) 9.
[61] D.-H. Lee, S.-J. Kim, Y.-W. Choi, Classification of Distraction Levels Using Hybrid Deep Neural Networks From EEG Signals, 11th International Winter Conference on Brain-Computer Interface (BCI), IEEE, 2023.
[62] Y. Li, K. Li, J. Chen, S. Wang, H. Lu, D. Wen, Pilot stress detection through physiological signals using a transformer-based deep learning model, IEEE Sens. J. 23 (2023) 11774–11784.
[63] Q. Li, K.K.H. Ng, C.Y. Yiu, X. Yuan, C.K. So, C.C. Ho, Securing air transportation safety through identifying pilot's risky VFR flying behaviours: an EEG-based neurophysiological modelling using machine learning algorithms, Reliab. Eng. Syst. Saf. 238 (2023) 15.
[64] J. Yuan, X. Ke, C. Zhang, Q. Zhang, C. Jiang, W. Cao, Recognition of different turning behaviors of pilots based on flight simulator and fNIRS data, IEEE Access 12 (2024) 32881–32893.
[65] Y. Wang, W.-C. Li, A. Nichanian, W.T. Korek, W.-T.-K. Chan, Future Flight Safety Monitoring: Comparison of Different Computational Methods for Predicting Pilot Performance Under Time Series During Descent by Flight Data and Eye-Tracking Data, 21st International Conference on Engineering Psychology and Cognitive Ergonomics, Springer, Washington, DC, 2024, pp. 308–320.
[66] B. Binias, D. Myszor, K.A. Cyran, A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals, Comput. Intell. Neurosci. 2018 (2018) 9.
[67] Z. Li, F. Li, M. Lyu, Tracking the Unseen and Unaware: Deciphering Controllers' Detection Failures to Warnings Through Eye-Tracking Metrics, Int. J. Hum.-Comput. Interact. (2025) 20.
[68] L. Shang, H. Si, H. Wang, T. Pan, H. Liu, Y. Li, J. Qiu, M. Xu, Research on fatigue detection of flight trainees based on face EMF feature model combination with PSO-CNN algorithm, Sci. Rep. 14 (2024) 11.
[69] Z. Huang, W. Tang, Q. Tian, T. Huang, J. Li, Air traffic controller fatigue detection based on facial and vocal features using long short-term memory, IEEE Access 12 (2024) 56663–56682.
[70] S. Luo, C. Zhang, W. Zhu, H. Chen, J. Yuan, Q. Li, T. Wang, C. Jiang, Noncontact perception for assessing pilot mental workload during the approach and landing under various weather conditions, Signal Image Video Process. 19 (2025) 17.
[71] S. Badrinath, H. Balakrishnan, Automatic speech recognition for air traffic control communications, Transp. Res Record 2676 (2022) 798–810.
[72] A. Arra, G. Achour, A.P. Payan, E.D. Harrison, D.N. Mavris, Automatic Speech Recognition Model Fine-Tuning and Development of a New Evaluation Metric for Terminal Airspace Safety Analysis, AIAA Aviation Forum, Amer Inst Aeronautics & Astronautics, Las Vegas, NV, 2024.
[73] Y. Lin, B. Yang, D. Guo, P. Fan, Towards multilingual end-to-end speech recognition for air traffic control, IET Intell. Transp. Syst. 15 (2021) 1203–1214.

[74] Y. Lin, B. Yang, L. Li, D. Guo, J. Zhang, H. Chen, Y. Zhang, ATCSpeechNet: a multilingual end-to-end speech recognition framework for air traffic control systems, Appl. Soft Comput. 112 (2021) 11.
[75] O. Ohneiser, U. Ahmed, Text-to-speech application for training of aviation radio telephony communication operators, IEEE Trans. Aerosp. Electron. Syst. 61 (2025) 4542–4560.
[76] K.L. Fox, K.R. Niewoehner, M. Rahmes, J. Wong, R. Razdan, Leverage Large Language Models for Enhanced Aviation Safety, IEEE, Herndon, VA, 2024.
[77] E. Pinska-Chauvin, H. Helmke, J. Dokic, P. Hartikainen, O. Ohneiser, R. G. Lasheras, Ensuring safety for artificial-intelligence-based automatic speech recognition in air traffic control environment, Aerospace 10 (2023) 23.
[78] Y. Lin, M. Ruan, K. Cai, D. Li, Z. Zeng, F. Li, B. Yang, Identifying and managing risks of AI-driven operations: a case study of automatic speech recognition for improving air traffic safety, Chin. J. Aeronaut. 36 (2023) 366–386.
[79] Y. Pang, J. Hu, C.S. Lieber, N.J. Cooke, Y. Liu, Air traffic controller workload level prediction using conformalized dynamical graph learning, Adv. Eng. Inform. 57 (2023) 16.
[80] N. Wu, J. Sun, Fatigue detection of air traffic controllers based on radiotelephony communications and self-adaption quantum genetic algorithm optimization ensemble learning, Appl. Sci. 12 (2022) 16.
[81] L. Wang, J. Chou, X. Zhou, A. Tien, D.M. Baumgartner, AviationGPT: A Large Language Model for the Aviation Domain, AIAA Aviation Forum, AIAA, Las Vegas, NV, 2024.
[82] V.M. Janakiraman, Explaining Aviation Safety Incidents Using Deep Temporal Multiple Instance Learning, 24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM, London, United Kingdom, 2018, pp. 406–415.
[83] X. Hu, J. Wu, J. He, Textual Indicator Extraction from Aviation Accident Reports, AIAA Aviation Forum and Exposition, AIAA, Dallas, TX, 2019.
[84] P. Srinivasan, V. Nagarajan, S. Mahadevan, Mining and Classifying Aviation Accident Reports, AIAA Aviation Forum and Exposition, AIAA, Dallas, TX, 2019.
[85] J.G. Fuller, L.R. Hook, Understanding General Aviation Accidents in Terms of Safety Systems, 39th AIAA/IEEE Digital Avionics Systems Conference (DASC), IEEE, 2020.
[86] Y. Gao, G. Zhu, Y. Duan, J. Mao, Semantic encoding algorithm for classification and retrieval of aviation safety reports, IEEE Trans. Autom. Sci. Eng. 8 (2024).
[87] D. Shi, J. Zurada, S. Cao, J. Guan, An Innovative Approach to Modeling Aviation Safety Incidents, in: 55th Annual Hawaii International Conference on System Sciences (HICSS), 2022, pp. 1216–1225.
[88] F.L. Lazaro, T. Madeira, R. Melicio, D. Valerio, L.F.F.M. Santos, Identifying human factors in aviation accidents with natural language processing and machine learning models, Aerospace 12 (2025) 20.
[89] T.T. Inan, N.G. Inan, The analysis of fatal aviation accidents more than 100 dead passengers: an application of machine learning, Opsearch 59 (2022) 1377–1395.
[90] X. Wang, Z. Gan, Y. Xu, B. Liu, T. Zheng, Extracting domain-specific chinese named entities for aviation safety reports: a case study, Appl. Sci. 13 (2023) 19.
[91] Y. Jiao, J. Dong, J. Han, H. Sun, Classification and causes identification of chinese civil aviation incident reports, Appl. Sci. 12 (2022) 19.
[92] E. Mangortey, A.H. Speirs, M.V. Bendarkar, V.P. Bui, Analysis of Helicopter Accidents and Certification Categories Using Machine Learning, AIAA, AIAA SciTech Forum and Exposition, 2022.
[93] J. Korentsides, J.R. Keebler, M. Berezovski, A. Chaparro, Factors contributing to fatalities in helicopter emergency medical service accidents, Aerosp. Med. Hum. Perform. 96 (2025) 119.
[94] X. Zhang, P. Srinivasan, S. Mahadevan, Sequential deep learning from NTSB reports for aviation safety prognosis, Saf. Sci. 142 (2021) 12.
[95] T. Dong, Q. Yang, N. Ebadi, X.R. Luo, P. Rad, Identifying incident causal factors to improve aviation transportation safety: proposing a deep learning approach, J. Adv. Transp. 2021 (2021) 15.
[96] G. Perboli, M. Gajetti, S. Fedorov, S. Lo Giudice, Natural Language Processing for the identification of Human factors in aviation accidents causes: an application to the SHEL methodology, Expert Syst. Appl. 186 (2021) 7.
[97] X. Ni, H. Wang, L. Chen, R. Lin, Classification of aviation incident causes using LGBM with improved cross-validation, J. Syst. Eng. Electron. 35 (2024) 396–405.
[98] A. Nanyonga, H. Wasswa, U. Turhan, O. Molloy, G. Wild, Sequential Classification of Aviation Safety Occurrences with Natural Language Processing, AIAA, AIAA Aviation Forum, 2023.
[99] A. Nanyonga, H. Wasswa, K. Joiner, U. Turhan, G. Wild, Explainable supervised learning models for aviation predictions in Australia, Aerospace 12 (2025) 21.
[100] M. Xiong, H. Wang, C. Che, R. Lin, Toward safer aviation: application of GA-XGBoost-SHAP for incident cognition and model explainability, Proc. Inst. Mech. Eng. Part O-J. Risk Reliab. 238 (2024) 1195–1208.
[101] Z. Zhuang, Y. Hou, L. Yang, J. Gong, L. Wang, Toward safer flight training: the data-driven modeling of accident risk network using text mining based on deep learning, Int. J. Comput. Intell. Syst. 17 (2024) 21.
[102] N. Niraula, S. Ayhan, B. Chidambaram, D. Whyatt, Multi-label Classification with Generative Large Language Models, 43rd AIAA DATC/IEEE Digital Avionics Systems Conference, IEEE, San Diego, CA, 2024.
[103] Y. Yang, D. Shi, J. Zurada, J. Guan, Application of Large Language Model and In-context Learning for Aviation Safety Prediction, 17th International Conference on Advanced Computer Theory and Engineering, IEEE, Hefei, China, 2024, pp. 361–365.
[104] V. Siddeshwar, A. Azim, S. Alwidian, M. Makrehchi, Towards Enhancing Aviation Safety through Advanced Incident Analysis using Large Language Models, 34th International Conference on Collaborative Advances in Software and Computing, IEEE, Toronto, Canada, 2024, pp. 221–227.
[105] L. Chen, J. Xu, T. Wu, J. Liu, Information extraction of aviation accident causation knowledge graph: an LLM-based approach, Electronics 13 (2024) 21.
[106] J. Emmons, T. Sharma, M. Salloum, B. Matthews, Text Summarization in Aviation Safety: A Comparative Study of Large Language Models, AIAA Aviation Forum, AIAA, Las Vegas, NV, 2024.
[107] A. Tikayat Ray, A.P. Bhat, R.T. White, V.M. Nguyen, O.J. Pinon Fischer, D. N. Mavris, Examining the potential of generative language models for aviation safety analysis: case study and insights using the aviation safety reporting system (ASRS), Aerospace 10 (2023).
[108] L. Irshad, H. Walsh, Identifying human errors and error mechanisms from accident reports using large language models, 2024 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference-IDETC-CIE, ASME, Washington, DC, 2024.
[109] M. Sujan, D. Slater, E. Crumpton, How can large language models assist with a FRAM analysis? Saf. Sci. 181 (2025) 10.
[110] Q. Liu, F. Li, K.K.H. Ng, J. Han, S. Feng, Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety, Expert Syst. Appl. 269 (2025) 126422.
[111] I.-V. Hernandez-Camero, E. Garcia-Lopez, A. Garcia-Cabot, S. Caro-Alvaro, Context-aware few-shot learning SPARQL query generation from natural language on an aviation knowledge graph, Mach. Learn. Knowl. Extr. 7 (2025) 18.
[112] T. Ren, Z. Zhang, B. Jia, S. Zhang, Retrieval-augmented generation-aided causal identification of aviation accidents: a large language model methodology, Expert Syst. Appl. 278 (2025) 18.
[113] F. Omrani, H. Etemadfard, R. Shad, Assessment of aviation accident datasets in severity prediction through machine learning, J. Air Transp. Manag. 115 (2024) 9.
[114] X. Zhang, S. Mahadevan, Ensemble machine learning models for aviation incident risk prediction, Decis. Support Syst. 116 (2019) 48–63.
[115] M. Caetano, Aviation accident and incident forecasting combining occurrence investigation and meteorological data using machine learning, Aviation 27 (2023) 47–56.
[116] H. Zeng, J. Guo, H. Zhang, B. Ren, J. Wu, Research on aviation safety prediction based on variable selection and LSTM, Sensors 23 (2023) 17.
[117] B. Cankaya, K. Topuz, A. Glassman, Business Inferences and Risk Modeling with Machine Learning; The case of Aviation Incidents, 56th Annual Hawaii International Conference on System Sciences (HICSS), Maui, HI, HICSS, 2023, pp. 1238–1248.
[118] D.V. Silagyi Ii, D. Liu, Prediction of severity of aviation landing accidents using support vector machine models, Accid. Anal. Prev. 187 (2023) 20.
[119] Y. Guo, Y. Sun, Y. He, F. Du, S. Su, C. Peng, Deep-learning-based model for accident-type prediction during approach and landing, IEEE Trans. Aerosp. Electron. Syst. 59 (2023) 472–482.
[120] A. Nanyonga, K. Joiner, U. Turhan, G. Wild, Natural language processing for aviation safety: predicting injury levels from incident reports in Australia, Modelling 6 (2025) 16.
[121] X. Ni, H. Wang, C. Che, J. Hong, Z. Sun, Civil aviation safety evaluation based on deep belief network and principal component analysis, Saf. Sci. 112 (2019) 90–95.
[122] R. Kaidi, M. Al Achhab, M. Lazaar, H. Omara, Improving the classification of airplane accidents severity using feature selection, extraction and machine learning models, Int. J. Adv. Comput. Sci. Appl. 14 (2023) 975–981.
[123] R.P.R. Nogueira, R. Melicio, D. Valerio, L.F.F.M. Santos, Learning methods and predictive modeling to identify failure by human factors in the aviation industry, Appl. Sci. 13 (2023) 15.
[124] T. Madeira, R. Melicio, D. Valerio, L. Santos, Machine learning and natural language processing for prediction of human factors in aviation incident reports, Aerospace 8 (2021) 18.
[125] M. Xiong, Z. Hou, H. Wang, C. Che, R. Luo, An aviation accidents prediction method based on MTCNN and Bayesian optimization, Knowl. Inf. Syst. 66 (2024) 6079–6100.
[126] S. Su, Y. Sun, C. Peng, Y. Guo, Improved gray correlation analysis and combined prediction model for aviation accidents, Eng. Comput. 40 (2023) 1570–1592.
[127] D. Zhou, X. Zhuang, H. Zuo, H. Wang, H. Yan, Deep learning-based approach for civil aircraft hazard identification and prediction, IEEE Access 8 (2020) 103665–103683.
[128] M. Xiong, H. Wang, Y.D. Wong, Z. Hou, Enhancing aviation safety and mitigating accidents: a study on aviation safety hazard identification, Adv. Eng. Inform. 62 (2024) 13.
[129] J. Ricketts, W. Guo, J. Pelham, D. Barry, Integrating an incident dataset with a question and answering language model to assist hazard identification: comparison of an extractive and generative model, Proc. Inst. Mech. Eng. Part O-J. Risk Reliab. 239 (2025) 736–753.
[130] Y. Su, Natural language processing system for text classification corpus based on machine learning, ACM Trans Asian Low-Resour. Lang. Inf. Process. 23 (2024) 15.
[131] Z. Hou, H. Wang, M. Xiong, C. Zhou, Y. Yue, Towards trustworthy civil aviation hazards identification: an uncertainty-aware deep learning framework, Adv. Eng. Inform. 65 (2025) 31.
[132] C. Chandra, X. Jing, M.V. Bendarkar, K. Sawant, L.R. Elias, M. Kirby, D.N. Mavri, Aviation-BERT: A Preliminary Aviation-specific Natural Language Model, AIAA, AIAA Aviation Forum, 2023.
[133] X. Jing, K. Sawant, M.V. Bendarkar, L.R. Elias, D.N. Mavris, Expanding Aviation Knowledge Graph using Deep Learning for Safety Analysis, AIAA Aviation Forum, AIAA, Las Vegas, NV, 2024.

[134] J. Oehling, D.J. Barry, Using machine learning methods in airline flight data monitoring to generate new operational safety knowledge from existing data, Saf. Sci. 114 (2019) 89–104.
[135] C. Tong, X. Yin, J. Li, T. Zhu, R. Lv, L. Sun, J.J.P.C. Rodrigues, An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data, Appl. Soft Comput. 73 (2018) 344–349.
[136] C. Guo, Y. Sun, T. Xu, Y. Hu, R. Yu, An improved transformer method for prediction of aircraft hard landing based on QAR data, Int. J. Aeronaut. Space Sci. 26 (2025) 2043–2057.
[137] D. Gil, A. Hernandez-Sabate, J. Enconniere, S. Asmayawati, P. Folch, J. Borrego-Carazo, M. Angel Piera, E-Pilots: a system to predict hard landing during the approach phase of commercial flights, IEEE Access 10 (2022) 7489–7503.
[138] E.V. Odisho, D. Truong, Applying machine learning to enhance runway safety through runway excursion risk mitigation. Integrated Communications Navigation and Surveillance Conference (ICNS), 2021.
[139] T.-Y. Chiu, Y.-C. Lai, Unstable approach detection and analysis based on energy management and a deep neural network, Aerospace 10 (2023) 25.
[140] Y. Kong, X. Zhang, S. Mahadevan, Bayesian deep learning for aircraft hard landing safety assessment, IEEE Trans. Intell. Transp. Syst. 23 (2022) 17062–17076.
[141] J. Cai, J. Shang, X. Li, C. Li, L. Zheng, Fine-grained time and hidden feature learning for interpretable hard landing prediction based on QAR data, IEEE Trans. Intell. Transp. Syst. 16 (2025).
[142] J. Shang, X. Li, R. Zhang, L. Zheng, X. Li, R. Zhang, X. Zhao, F. Li, H. Sun, A dual two-stage attention-based model for interpretable hard landing prediction from flight data, Eng. Appl. Artif. Intell. 154 (2025) 12.
[143] C. Tong, X. Yin, S. Wang, Z. Zheng, A novel deep learning method for aircraft landing speed prediction based on cloud-based sensor data, Futur. Gener. Comp. Syst. 88 (2018) 552–558.
[144] Z. Kang, J. Shang, Y. Feng, L. Zheng, D. Liu, B. Qiang, R. Wei, A Deep Sequence-to-Sequence Method for Aircraft Landing Speed Prediction Based on QAR Data, in: 21st International Conference on Web Information Systems Engineering (WISE), 2020, pp. 516–530.
[145] H. Chen, J. Shang, X. Zhao, X. Li, L. Zheng, F. Chen, A Deep Learning Method for Landing Pitch Prediction based on Flight Data, 2nd IEEE International Conference on Civil Aviation Safety and Information Technology (ICCASIT), IEEE, Wuhan, China, 2020, pp. 199–204.
[146] Z. Kang, J. Shang, Y. Feng, L. Zheng, Q. Wang, H. Sun, B. Qiang, Z. Liu, A deep sequence-to-sequence method for accurate long landing prediction based on flight data, IET Intell. Transp. Syst. 15 (2021) 1028–1042.
[147] T.G. Puranik, N. Rodriguez, D.N. Mavris, Towards online prediction of safety-critical landing metrics in aviation using supervised machine learning, Transp. Res. Pt. C-Emerg. Technol. 120 (2020) 18.
[148] M. Memarzadeh, A.A. Asanjan, B. Matthews, Robust and explainable semi-supervised deep learning model for anomaly detection in aviation, Aerospace 9 (2022) 21.
[149] A. Nichanian, D. Koch, W.C. Li, Applying artificial neural networks for multidimensional anomaly detection based on flight data monitoring during final approaches, Aeronaut. J. 18 (2025).
[150] A. Khattak, J. Zhang, P.-W. Chan, F. Chen, C.M. Matara, AI-supported estimation of safety critical wind shear-induced aircraft go-around events utilizing pilot reports, Heliyon 10 (2024) 19.
[151] X. Wang, R. Mou, Flight safety risk prediction for civil aircraft approach and landing, J. Aerosp. Inf. Syst. 22 (2025) 220–230.
[152] P.-C. Tsai, Y.-C. Lai, Risk assessment procedure of final approach to landing using deep learning, J. Aerosp. Inf. Syst. 21 (2024) 323–331.
[153] L.C.E. Silva, M.C.R. Murca, Machine learning models for online anomaly detection in flight operations, AIAA, AIAA Aviation Forum, 2023.
[154] J.J.M. Lopetegui, M. Tanelli, Combining model-based and learning-based anomaly detection schemes for increased performance and safety of aircraft braking controllers, Eng. Appl. Artif. Intell. 139 (2025) 18.
[155] K. Dmitriev, J. Rhein, L. Beller, J. Broecker, E. Huber, J. Schumann, F. Holzapfel, Safety Assessment of a Machine Learning-based Aircraft Emergency Braking System: A Case Study, 43rd AIAA DATC/IEEE Digital Avionics Systems Conference, IEEE, San Diego, CA, 2024.
[156] P.C. Mural, G.N. Rathna, V. Bhola, Active Learning in Flight Anomaly Detection, AIAA SciTech Forum, AIAA, Orlando, FL, 2024.
[157] L. Gao, C. Xu, F. Wang, J. Wu, H. Su, Flight data outlier detection by constrained LSTM-autoencoder, Wirel. Netw 29 (2023) 3051–3061.
[158] E. Mangortey, D. Monteiro, J. Ackley, Z. Gao, T.G. Puranik, M. Kirby, O.J. Pinon, D.N. Mavri, Application of Machine Learning Techniques to Parameter Selection for Flight Risk Identification, AIAA SciTech Forum and Exposition, AIAA, Orlando, FL, 2020.
[159] A. Tato, R. Nkambou, G.J.N. Tato, Towards Adaptive Coaching in Piloting Tasks: Learning Pilots' Behavioral Profiles from Flight Data, 18th International Conference on Intelligent Tutoring Systems (ITS), Springer, Bucharest, Romania, 2022, pp. 105–114.
[160] M. Xiong, H. Wang, Z. Hou, Y.D. Wong, Multi-level information identification for civil aviation safety risks: a hierarchical multi-branch deep learning approach, Inf. Sci. 702 (2025) 20.
[161] H. Sun, F. Yang, P. Zhang, Y. Jiao, Y. Zhao, An innovative deep architecture for flight safety risk assessment based on time series data, CMES-Comp. Model. Eng. Sci. 138 (2024) 21.
[162] J. Li, H. Zhang, J. Yang, Discrimination Model of QAR High-Severity Events Using Machine Learning, 9th International Conference on Intelligence Science and Big Data Engineering (IScIDE), Springer, Nanjing, China, 2019, pp. 430–441.
[163] M.A. Rahman, T. Bhuiyan, M.A. Ali, Enhancing aviation safety: machine learning for real-time ADS-B injection detection through advanced data analysis, Alex. Eng. J. 126 (2025) 262–276.
[164] T.G. Puranik, D.N. Mavris, Identification of instantaneous anomalies in general aviation operations using energy metrics, J. Aerosp. Inf. Syst. 17 (2020) 51–65.
[165] A. Grushin, J. Nanda, A. Tyagi, D. Miller, J. Gluck, N.C. Oza, A. Maheshwari, Decoding the Black Box: Extracting Explainable Decision Boundary Approximations from Machine Learning Models for Real Time Safety Assurance of the National Airspace, AIAA SciTech Forum and Exposition, AIAA, San Diego, CA, 2019.
[166] M. Memarzadeh, B. Matthews, T. Templin, Multiclass anomaly detection in flight data using semi-supervised explainable deep learning model, J. Aerosp Inf. Syst. (2021) 15.
[167] M. Memarzadeh, B. Matthew, T. Templin, Multi-Class Anomaly Detection in Flight Data Using Semi-Supervised Explainable Deep Learning Model, AIAA, AIAA SciTech Forum and Exposition, 2021.
[168] X. Li, J. Shang, L. Zheng, Q. Wang, D. Liu, X. Liu, F. Li, W. Cao, H. Sun, IMTCN: An Interpretable Flight Safety Analysis and Prediction Model Based on Multi-Scale Temporal Convolutional Networks, IEEE Trans. Intell. Transp. Syst. 25 (2024) 289–302.
[169] H. Lee, S. Madar, S. Sairam, T.G. Puranik, A.P. Payan, M. Kirby, O.J. Pinon, D. N. Mavris, Critical parameter identification for safety events in commercial aviation using machine learning, Aerospace 7 (2020) 24.
[170] Z. Xiang, Z. Gao, Y. Gao, Y. Zhang, R. Zhang, Real-time identification of precursors in commercial aviation using multiple-instance learning, Adv. Eng. Inform. 62 (2024) 12.
[171] M.A. Ramos, K. Sankaran, S. Guarro, A. Mosleh, R. Ramezani, A. Arjounilla, The need for and conceptual design of an AI model-based Integrated Flight Advisory System, Proc. Inst. Mech. Eng. Part O-J. Risk Reliab. 237 (2023) 485–507.
[172] Air Line Pilots Association International, White Paper: The Dangers of Single-Pilot Operations, 2019.
[173] Chartered Institute of Ergonomics & Human Factors, White Paper: The Human Dimension in Tomorrow's Aviation System, 2020.
[174] L. Dong, H. Chen, C. Zhao, P. Wang, Analysis of single-pilot intention modeling in commercial aviation, Int. J. Aerosp. Eng. 2023 (2023) 14.
[175] F. Li, S. Feng, Y. Yan, C.-H. Lee, Y.S. Ong, Virtual Co-pilot: Multimodal Large Language Model-enabled Quick-access Procedures for Single Pilot Operations, 2nd IEEE Conference on Artificial Intelligence (CAI), IEEE, Singapore, Singapore, 2024, pp. 1501–1506.
[176] S. Wen, M. Middleton, S. Ping, N.N. Chawla, G. Wu, B.S. Feest, C. Nadri, Y. Liu, D. Kaber, M. Zahabi, R.P. McMahan, S. Castelo, R. McKendrick, J. Qian, C. T. Silva, AdaptiveCoPilot: Design and Testing of a NeuroAdaptive LLM Cockpit Guidance System in both Novice and Expert Pilots, 32nd Conference on Virtual Reality and 3D User Interfaces-VR, IEEE, Saint Malo, France, 2025, pp. 656–666.
[177] Y. Guo, Y. Sun, Y. He, F. Du, S. Su, C. Peng, A data-driven integrated safety risk warning model based on deep learning for civil aircraft, IEEE Trans. Aerosp. Electron. Syst. 59 (2023) 1707–1719.
[178] Y. Han, X. Huang, Autonomous air traffic separation assurance through machine learning, J. Ind. Manag. Optim. 20 (2024) 3195–3204.
[179] V. Janson, A. Ahlbrecht, U. Durak, Architectural Challenges in Developing an AI-based Collision Avoidance System, IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC), IEEE, Barcelona, Spain, 2023.
[180] G. Papadopoulos, A. Bastas, G.A. Vouros, I. Crook, N. Andrienko, G. Andrienko, J. M. Cordero, Deep reinforcement learning in service of air traffic controllers to resolve tactical conflicts, Expert Syst. Appl. 236 (2024) 19.
[181] P. Ortner, R. Steinhoefler, E. Leitgeb, H. Fluehr, Augmented air traffic control system-artificial intelligence as digital assistance system to predict air traffic conflicts, AI 3 (2022) 623–644.
[182] T. Stefani, M. Jameel, I. Gerdes, R. Hunger, C. Bruder, E. Hoemann, J. M. Christensen, A.A. Girija, F. Koester, T. Krueger, S. Hallerbach, Towards an Operational Design Domain for Safe Human-AI Teaming in the Field of AI-Based Air Traffic Controller Operations, 43rd AIAA DATC/IEEE Digital Avionics Systems Conference, IEEE, San Diego, CA, 2024.
[183] D.-T. Pham, S. Alam, V. Duong, An air traffic controller action extraction-prediction model using machine learning approach, Complexity 2020 (2020) 19.
[184] Y. Lin, L. Deng, Z. Chen, X. Wu, J. Zhang, B. Yang, A real-time ATC safety monitoring framework using a deep learning approach, IEEE Trans. Intell. Transp. Syst. 21 (2020) 4572–4581.
[185] K. Van Benthem, C.M. Herdman, A virtual reality cognitive health screening tool for aviation: managing accident risk for older pilots, Int. J. Ind. Ergon. 85 (2021) 12.
[186] Y. Meng, Z. Peng, Z. Zhang, Q. Chen, H. Huang, Y. Chen, M. Zhao, Predicting honest behavior based on Eysenck personality traits and gender: an explainable machine learning study using SHAP analysis, Front. Psychol. 16 (2025) 18.
[187] J. Zhang, P. Zhang, D. Guo, Y. Zhou, Y. Wu, B. Yang, Y. Lin, Automatic repetition instruction generation for air traffic control training using multi-task learning with an improved copy network, Knowledge-Based Syst. 241 (2022) 12.
[188] Y. Lin, Y. Wu, D. Guo, P. Zhang, C. Yin, B. Yang, J. Zhang, A deep learning framework of autonomous pilot agent for air traffic controller training, IEEE T Hum.-Mach. Syst. 51 (2021) 442–450.
[189] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–215.
[190] A. Schepman, P. Rodway, Initial validation of the general attitudes towards Artificial Intelligence Scale, Comput. Hum. Behav. Rep. 1 (2020) 100014.

[191] J. Dong, S. Chen, M. Miralinaghi, T. Chen, P. Li, S. Labi, Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems, Transp. Res. Part C Emerging Technol. 156 (2023) 104358.
[192] M.R. Endsley, Supporting human-AI teams:transparency, explainability, and situation awareness, Comput. Hum. Behav. 140 (2023) 107574.
[193] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, R. Ranjan, Explainable AI (XAI): core ideas, techniques, and solutions, ACM Comput. Surv. 55 (2023). Article 194.
[194] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 4768–4777.
[195] EASA, Artificial Inteliigence Roadmap 2.0: Human-centric approach to AI in aviation, 2023.
[196] EU High-Level Expert Group on AI, Ethics guidelines for trustworthy AI, 2019.
[197] EASA, EASA's first regulatory proposal on Artificial Intelligence for Aviation is now open for consultation, 2025.
[198] V. Battiste, J. Lachter, S. Brandt, A. Alvarez, T.Z. Strybel, K.-P.-L. Vu, Human-Automation Teaming: Lessons Learned and Future Directions, in: S. Yamamoto, H. Mori (Eds.), Human Interface and the Management of Information, Springer International Publishing, Cham, Information in Applications and Services, 2018, pp. 479–493.
[199] J. Wuerfel, A. Papenfus, M. Wies, Operationalizing AI Explainability Using Interpretability Cues in the Cockpit: Insights from User-Centered Development of the Intelligent Pilot Advisory System (IPAS), 26th International Conference on Human-Computer Interaction (HCII), Springer, Washington, DC, 2024, pp. 297–315.
[200] A. Saraf, K. Chan, M. Popish, J. Browder, J. Schade, Explainable Artificial Intelligence for Aviation Safety Applications, AIAA, AIAA Aviation Forum, 2020.

Enhancing aviation safety with artificial intelligence: a systematic literature review on recent advances, challenges and future perspectives

Yiu, Cho Yin

2026-04

Yiu CY, Li W-C, Ng KKH, et al., (2026) Enhancing aviation safety with artificial intelligence: a systematic literature review on recent advances, challenges and future perspectives. Advanced Engineering Informatics, Volume 71, Part B, April 2026, Article number 104378

https://doi.org/10.1016/j.aei.2026.104378