Highlights

Tracking the Unseen and Unaware: Deciphering Controllers' Detection Failures to Warnings through Eye-tracking Metrics

Zhimin Li, Fan Li, Mengtao Lyu

- Detection failure classification targets different situation awareness levels.
- Distinct gaze patterns mark different detection failure types.
- Random forest leads with 80% precision in a four-detection-type recognition task.
- Continuous warnings weaken the awareness of warnings but enhance projection.

Tracking the Unseen and Unaware: Deciphering Controllers' Detection Failures to Warnings through Eye-tracking Metrics

Zhimin Li[a], Fan Li[a,*], Mengtao Lyu[a]

[a]*Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong, China*

Abstract

With the integration of digital towers in air traffic control (ATC), the increasing complexity of visual cues poses challenges in detecting vital warning signals by controllers. The detection failure (DF) to warnings leads to unawareness of potential airspace hazards, making its timely identification crucial. However, the inherent variability in human situational awareness and behaviors complicates the differentiation and recognition of various DFs. This paper aims to decipher DF by categorizing it into different types based on Endsley's situation awareness theory and identifying the specific causes and key indicators of each type. A four-phase framework is proposed, including gaze-behavioral DF classification, DF induction experiment, gaze dynamics across various DFs and warning frequencies, and DF-type recognition and analytics. Utilizing this framework, gaze data from 26 subjects were collected via a DF induction experiment. Analysis of these data revealed that non-perception, unaware perception, and aware perception of warnings are observed in markedly different gaze patterns. Furthermore, continuous warnings primarily weaken operators' awareness of warnings while improving the foresight of warning implications. Additionally, DF-type recognition and analytics were conducted, with the random forest model achieving up to 80% precision in four-category recognition. This research provides empirical evidence for real-time DF-type recognition and targeted intervention developments, offering crucial insights for the enhancement of visual warning detection and effective human-computer interaction in aviation safety.

Keywords: Warning detection, Detection failures recognition, Eye tracking, Warning frequency, Air traffic control

*Corresponding author.

Email addresses: zhimin.li@connect.polyu.hk (Zhimin Li), fan-5.li@polyu.edu.hk (Fan Li), mengtao.lyu@connect.polyu.hk (Mengtao Lyu)

1. Introduction

Numerous air traffic control (ATC) incidents arise from overlooking potential safety risks due to operators' focused awareness being limited to only a small subset of the visual environment (Varakin et al., 2004). An automated early warning system, communicating warnings of varying urgency levels to air traffic controllers (ATCOs) via a human-computer interface, is therefore indispensable (Wu & Li, 2018). There is a crucial requirement for humans operating in ATC: the capacity to responsibly process meaningful information from an automated warning system, even in the absence of immediate action (Eißfeldt et al., 2010). However, while the control support tools may reduce the likelihood of human errors, the introduction of new tasks could also bring new sources of cognitive errors (Corver & Aneziris, 2015). For instance, there are excessive warnings in the ATC domain, as missing is normally much more terrible than false alarms or unnecessary alarms in ATC. Evidence from the terminal environment indicates that 44% of conflict alerts and 61% of minimum safe altitude warnings were assessed as not necessitating intervention from ATCOs (Friedman-Berg et al., 2008). Under the situation of excessive warnings, operators might overlook or deprioritize the unexpected warnings, leading to a heightened risk of detection failure (DF) to the warnings (Liu et al., 2020).

DF, a type of human error originating from cognitive psychology, is characterized by the failure to accurately detect unexpected but salient signals, such as warnings (Simons & Chabris, 1999). It represents 50% of failures of safety risk perception for hazards with accident potential (Park et al., 2022). The recent Haneda Airport accident in Japan, where ATCOs missed a critical blinking warning signal leading to a runway collision, exemplifies the urgent need for enhanced DF management in ATC (Nikkei, 2024). Given the severe implications of DF, considerable attention has been paid to DF recognition, intervention design, and reduction(Bruder & Hasse, 2020; Saward & Stanton, 2018; Werneke & Vollrath, 2011). Recognizing DF is a crucial initial step, serving as a reactive strategy to alert ATCOs about ignored or misunderstood warnings. Kontogiannis & Malakis (2009) proposed a cognitive strategy framework for DF detection through two detection mechanisms, namely awareness-based detection and planning-based detection. Nevertheless, traditional DF identification methods, like self-reporting and retrospective analysis, are subjectively limited, as noted by Mack & Rock (1998). Recent studies have shifted to objectively assessing physiological aspects such as eye movements and brain activities during monitoring (Friedrich et al., 2017; Li et al., 2020; Lyu et al., 2024a). For instance, research has examined the extent of visual processing and consciousness in scenarios

lacking attention and awareness, as well as the correlation between visual parameters and ATCOs' monitoring behaviors related to automation failures (Hutchinson, 2019; Bruder & Hasse, 2019). However, few studies have focused on the importance of the DF concept and its recognition using the eye-tracking technique in the ATC industry.

In this study, we aim to reveal the effective DF indicators and recognize DF through the eye-tracking technique. This idea is supported by prior studies that collectively underscored eye-tracking as a non-intrusive tool for assessing operators' visual attention (Li et al., 2022). Vision is a vital channel for human-computer interactions in cognitive tasks, especially in supervisory functions. Varakin et al. (2004) discussed the pivotal role of visual awareness in human-computer interface design, highlighting how the limitations in our perception can lead to significant DFs. Metrics like fixations and pupil responses are identified as reliable indicators of understanding attention shifts and cognitive load (Mengtao et al., 2023). Saccades relates to scanning strategies (Bodala et al., 2016). Besides, Li et al. (2024) demonstrated the notable eye movement patterns linked to the inattentional blindness phenomenon, which is a detection failure to an unexpected but salient stimulus. Renata et al. (2018) also confirmed that the relations between saccade peak velocity and the first view time are affected by cognitive states.

Nevertheless, to achieve gaze-based DF recognition, several challenges have to be addressed. First, few studies have classified DF systematically. DF is associated with the complex interaction between human situational awareness and decision-making behaviors, classifying it aids in understanding its nature, causes, and potential impact, enabling targeted mitigation strategies. Based on Endsley's situation awareness theory, warning detection involves three phases: perception (perceiving a warning), comprehension (understanding the warning), and projection (foreseeing potential risks) (Endsley, 1995; Wu et al., 2023). Disruption in the three stages would induce different DF types, each necessitating specific countermeasures. Specifically, it is termed as ordinary blindness (OB) when warnings are completely overlooked, indicating a need for improved displays (Ruskin et al., 2021). Look but fail to see (LBFTS) error, caused by gaps in the awareness of the perceived warnings, has no response and suggests reducing cognitive load (Wang et al., 2022). Misinterpretation (MI) condition involves processing but misunderstanding warnings, causing erroneous foresight of warnings and thus necessitating enhanced training (Bruder & Hasse, 2020). Hence, distinguishing these DF types is crucial for developing targeted interventions. Second, the warning frequency would affect gaze movements and thus should be considered during data collection, as evidenced by increased missed warnings among ATCOs during excessive warning periods (Ruskin et al., 2021). Third,

identifying suitable gaze-based DF indicators is challenging due to the inherent connection between gaze patterns and the detection process. Fixations and saccades, essential to visual search and decision-making, often overlap with successful detection, making it difficult to distinguish between detection failures and routine visual processing (Holmqvist et al., 2011). Additionally, gaze patterns related to DF are subtle and vary across individuals, further complicating the identification of consistent indicators (Orquin & Loose, 2013). Finally, the complex interplay between DF and eye movements necessitates the careful selection of an optimal machine-learning method from a wide array and an in-depth analysis of feature significance. Hence, recently widely applied machine learning and feature importance methods may be utilized for recognition (Zhou et al., 2021; Lyu et al., 2024b).

To address the above challenges, a four-phase framework is proposed, including gaze-behavioral DF classification, DF induction experiment, gaze dynamics across various DFs and warning frequencies, and DF-type recognition and analytics. In the initial phase, we define detection types based on subjects' visual and behavioral responses, classifying them into correct detection and three DF types. Following this, a DF induction experiment simulates basic ATCOs' tasks by requiring participants to monitor aircraft (Bruder & Hasse, 2020) and respond to two types of warning frequencies, continuous and interval warnings. The third phase involves gaze pattern analysis across DF types and the impact of warning frequency on DF types reflected by gaze patterns. In this phase, we test the following two hypotheses: H1: different DF types have significantly different gaze patterns. H2: warning frequency can significantly affect the gaze patterns under distinct DF types. The final phase applies three basic and four advanced tree-based machine learning models for detection type recognition. Upon identifying the optimal model, Shapley Additive Explanations (SHAP) are employed for model interpretation and feature importance analysis (Zhou et al., 2021). In this phase, we test the following hypotheses: H3: the gaze features and warning frequency can be used to recognize four detection types.

This study establishes a basis for DF classification and DF-type recognition through eye movements, enriching our understanding of attentional and cognitive dynamics in ATC. It also guides the development of targeted strategies to mitigate these DF-induced issues in ATC and similar monitoring-intensive fields. Moreover, by understanding how warning frequency influences DF-type recognition, our findings provide scientific references for ATC management in automatic warning system design, thereby enhancing visual warning detection, and ultimately improving human-computer interaction in aviation.

2. Related works

2.1. DF measurements with eye movements

The eye movement-based measurements of DF in ATC are critical for DF recognition. Eye-tracking technology enables precise tracking of various eye movements, providing insights into how controllers monitor, process, and respond to visual stimuli (Castritius et al., 2021; Mengtao et al., 2023). Studies have identified key eye movement metrics as reliable indicators of visual attention shifts and cognitive load, such as patterns of fixations, saccadic movements, and pupillary parameters (Li et al., 2023). Specifically, fixation count and fixation duration are among the most commonly used metrics. Fixation count reflects the number of times the eye pauses within a particular area of interest (AOI), which correlates with attention and information processing (Goldberg & Kotval, 1998), while fixation duration, the length of these pauses, has been linked to the depth of cognitive processing(Bodala et al., 2017; Rayner, 1998). When automation failures occurred, operators were observed to fixate most frequently on the relevant AOIs (Bruder & Hasse, 2020). Richards et al. (2012) found that inattentional blindness (IB) was associated with making more fixations and longer gaze times on distractor stimuli, being less likely to fixate the unexpected stimulus. IB is a phenomenon that represents no fixations on the obvious but unexpected stimulus or fixations without attention to stimulus, encompassing OB and LBFTS (Liao & Chiang, 2016; Hou et al., 2024). Besides, Ball et al. (2014) offered a guide on using natural scenes to explore inattentional detection failure via image manipulation for change blindness studies. Mean saccade amplitude is indicative of the scanning strategy and information intake (Bodala et al., 2016). Moreover, studies in eye-tracking have allowed for more measures, such as mean pupil diameter, which is sensitive to changes in cognitive effort. For instance, Moacdieh et al. (2023) demonstrated that pupil size would be larger when DF occurs compared to normal conditions. In addition to the key eye-tracking features, many scholars refer to the first view time to the stimuli when DF occurs. For example, Renata et al. (2018) investigated the first view time in responding to the changes under different fatigue levels. Ruscio et al. (2015) quantified the reaction time of brakes to evaluate drivers' response to unexpected events. Although the above studies have not distinguished specific DF types, the conclusions of the above indicators help understand the visual response, attention distribution, and cognitive workload of operators when DF occurs.

2.2. Effects of warning frequency on DF

Prior research has extensively explored the factors influencing DF from various perspectives, including working memory capacity, task load, perceptual differences,

and habituation(Anderson et al., 2016; Brinton Anderson et al., 2016). For example, DF has been linked to limited working memory and high cognitive load, while individual perception differences have also been identified as contributing factors to DF (Richards et al., 2012; Li et al., 2023; Simons & Jensen, 2009). Additionally, the impact of warning frequency on alert effectiveness has been evaluated using a distributed signal detection theory model to assess continuous versus selective warning strategies (Papastavrou & Lehto, 1995). Frequent warnings have been linked to information overload, redundancy, and warning fatigue (Akhawe & Felt, 2013). Ruscio et al. (2015) emphasized the impact of warning information completeness on perception, evaluation, and decision-making in control settings. Additionally, the presence of excessive warnings, especially when warnings are deemed unreliable, can trigger the "cry wolf" effect, where operators may disable or deprioritize alarms (Papastavrou & Lehto, 1996). In situations of excessive warnings, there can be a discrepancy between the alarm's intended function and the user's perception of it, potentially eroding trust (Dixon et al., 2007; Jones et al., 2021). Recent findings from a study analyzing reports from the Aviation Safety Reporting System (ASRS) by Ruskin et al. (2021) showed that ATCOs are more likely to miss warnings during periods of excessive alarms. Navarro et al. (2019) noted that a missed warning hinders driving performance, while a false warning affects performance in subsequent lane departures. Besides, research on warning systems demonstrated that two-stage warning systems, HUD-based warnings, and peripheral signal designs each enhanced operator situational awareness and performance in driving and high-stakes environments (Ma et al., 2021; Yang et al., 2019; Werneke & Vollrath, 2011). The above-emerging evidence underscores the need for further research integrating eye metrics and different DF types to better understand the challenges posed by warning frequency.

3. Methods

In this study, we aim to systematically classify DF and identify the specific causes and key indicators of each type through an experimental and analytical framework. The framework, depicted in Figure 1, outlines the four-phase experimental and analytical process, illustrating the flow from DF classification and experiment setup to final recognition and analysis. The subsequent sections will detail each phase of this framework. To establish a clear foundation for this process, we formalize the study's objectives mathematically. Let $x_{i,e} \in \mathbb{R}^n$ represent the eye-tracking features of participant i for event e, with $y_{i,e} \in \{OB, LBFTS, MI, CD\}$ denoting the corresponding DF type label. The classification model f is optimized to capture the differences between DF types and select the best-performing model by maximizing

evaluation metrics including accuracy, precision, recall, and F1-score. This can be expressed as:

$$f^* = \arg\max_f \sum_{i=1}^{N} \sum_{e=1}^{E_i} M(f(x_{i,e}), y_{i,e})$$

where M represents the evaluation metrics. The distinct gaze patterns among the DF types are further validated through statistical analysis and SHAP-based feature importance evaluation, ensuring a clear differentiation between these DF types.

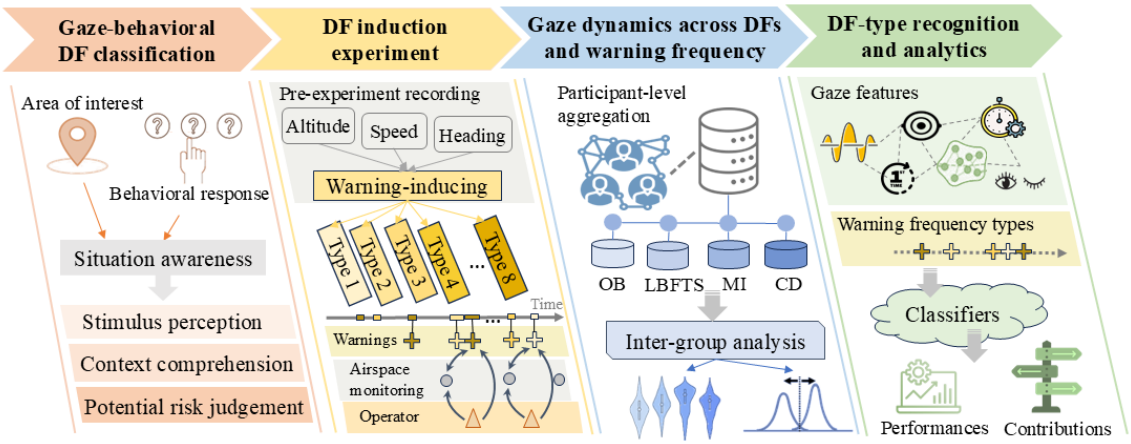


Figure 1: Flowchart illustrating the four-phase framework for DF classification and recognition.

3.1. Gaze-behavioral DF classification

Endsley's situation awareness theory outlines a three-phase detection process – perception, comprehension, and projection – which aligns with the typical stages where DF occurs in complex and dynamic environments like aviation (Endsley, 1995). Specifically, perception refers to the detection of a stimulus in the environment, comprehension involves the awareness of this stimulus to understand the current situation, and projection is the ability to foresee future states of the environment based on the understanding of the stimulus. These stages collectively contribute to effective detection and decision-making, particularly in dynamic contexts where rapid response to changes is crucial (Han et al., 2024). As shown in Figure 2, according to the three-phase detection process, DF can be categorized into three types through gaze and behavioral responses.Specifically, when an unexpected warning appears, instances, where subjects concentrate on a single task and fail to visually fixate on the warnings, are categorized as ordinary blindness (OB) to warnings (Hollnagel, 2000).

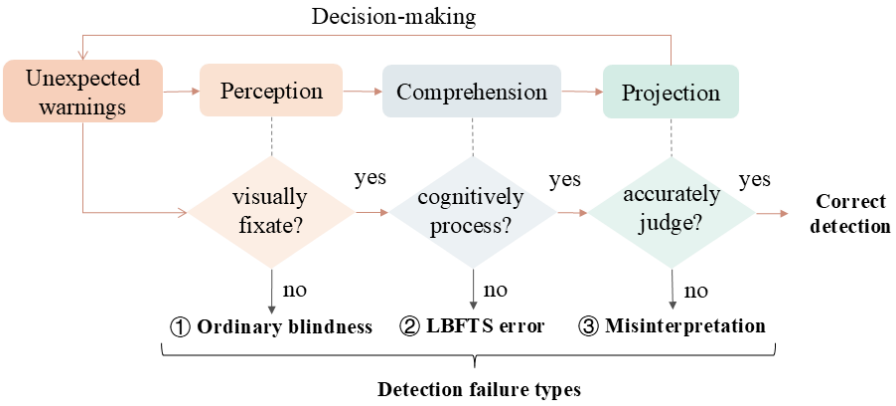


Figure 2: The proposed gaze-behavioral DF classification based on Endsley's situation awareness theory.

Subsequently, instances, where an individual perceives warnings but fails to cognitively process and comprehend them are classified as look-but-fail-to-see (LBFTS) errors (Wang et al., 2022). It indicates that the subjects perceive the new warning, but they are unaware of it and thus do not cognitively comprehend it. Since comprehension is a covert cognitive activity and difficult to be observed directly, we infer its occurrence through participant responses with the explicit instructions that require subjects to report any warning perceived (Xu et al., 2021). Furthermore, if subjects fail to accurately judge the warning's attributes or potential consequences, leading to flawed risk projections about future events, this is defined as misinterpretation (MI). This is consistent with Sarter & Woods (2017), who highlight that misinterpretation of critical cues in dynamic environments often leads to faulty situation awareness and inaccurate predictions of future events. Finally, if the subject perceives and correctly records the warnings, this is defined as correct detection (CD). Due to different circumstances and attribution, clearly distinguishing DF types in studying DF indicators and DF mechanisms is necessary and valuable for developing efficient interventions accordingly.

3.2. DF induction experiment

3.2.1. Participants

An experiment was conducted at the Hong Kong Polytechnic University to collect data and demonstrate the proposed framework. The research protocol underwent rigorous review and received approval (HSEARS20211117002), adhering to the institution's ethical standards and guidelines. Twenty-six postgraduate students from The

Hong Kong Polytechnic University, aged 20 to 30 years (M=25.65, SD=2.69), were recruited. They were selected based on their training in basic airspace monitoring and warning recognition, ensuring they had the necessary task-related skills. As the experiment involved simplified and abstract interface monitoring tasks, participants without prior experience as traffic controllers were appropriate for the study (Bruder & Hasse, 2020). All participants had either normal or corrected-to-normal vision. Informed consent was obtained from all participants prior to the commencement of the experiment.

3.2.2. Apparatus

In the methodological framework of this study, a desktop computer with a 27-inch monitor (1920*1080 pixels) was paired with a Gazepoint 3 (GP3) eye-tracking apparatus to gather intricate eye movement metrics, as depicted in Figure 3. The eye tracker software starts and runs with an extra laptop, which is connected to the monitor by HDMI. The GP3 is a desktop-integrated eye-tracking system, capable of recording a diverse range of eye movement metrics at a temporal resolution of 60 Hz. It targets the subject's eyes from an upward direction, ideally positioned 30 cm below the eye level and at a distance of 65 cm.

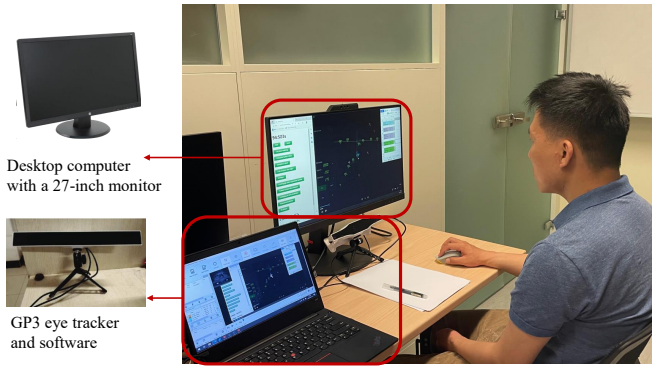


Figure 3: Experiment setting.

3.2.3. Experiment procedures

The research was conducted through systematically organized procedures to ensure accurate and reliable results, including five phases: introduction, training, practice, break and eye tracker calibration, and formal experiment, as shown in Figure 4.

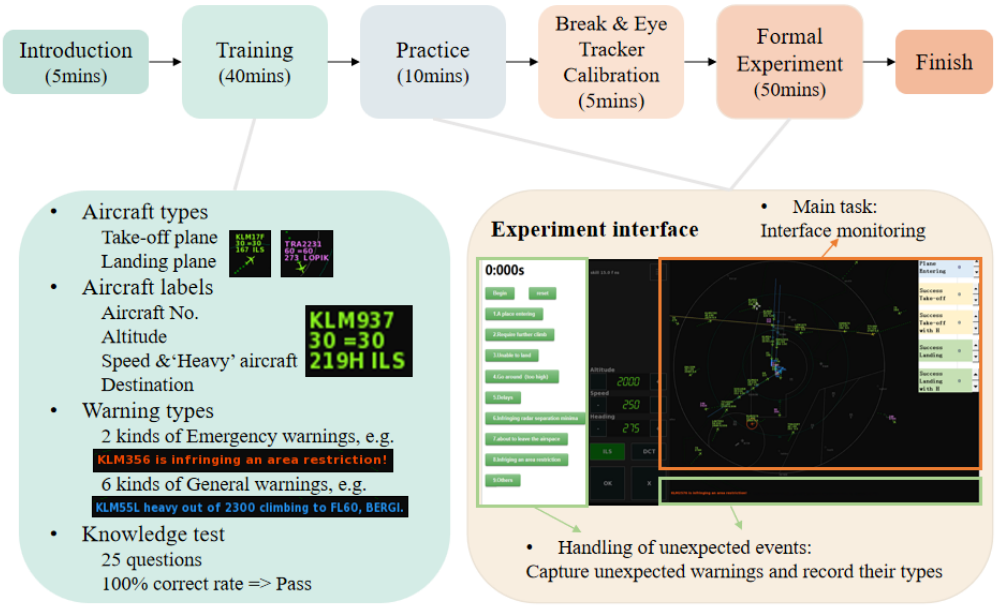


Figure 4: Experiment procedures.

Introduction. Participants are initially introduced to the overarching objectives and methodologies of the research. This foundational step ensures a uniform baseline understanding of the study's context across participants.

Training. The training process consisted of two key stages: familiarization with aircraft types and labels, and identification of various warning types. In the 40-minute training phase, subjects receive instruction on the types of aircraft in the Endless ATC platform, a realistic and interactive air traffic control simulator designed to provide a user-friendly environment for simulating airspace management tasks, as described in Yu et al. (2023). It allows users to manage multiple flights in a dynamic airspace by providing functionalities such as radar separation monitoring, altitude control, and direction adjustments. The platform's capability to generate complex scenarios and warnings makes it ideal for studying detection failures and human response patterns in a controlled environment.

In the 40-minute training phase, participants first learned to differentiate between two main types of aircraft: take-off planes with purple labels (leaving the airspace) and landing planes with green labels (entering the airspace). They were also familiarized with key flight label parameters, such as aircraft number, altitude, speed, destination, and whether it was classified as heavy (denoted by an 'H' indicating

increased separation during landing). Next, participants were trained to recognize 8 types of warnings. Two emergency warnings (radar separation and area restriction breaches) appeared in red text, while six general warnings (airspace entry, climb requirements, landing issues, detours, delays, and airspace exit) were displayed in blue text. To evaluate their understanding, participants completed a 25-question warning identification test, where each question presented a randomly ordered warning scenario, and participants had to identify its type correctly. All 25 questions needed to be answered correctly before advancing to the experimental phase. To further clarify, we have included a table in the appendix (Table A1) that outlines the warning examples corresponding to questions used in the test.

Practice. After training, participants had a short practice phase. They engaged in tasks similar to those in the formal experiment phase, ensuring that they were ready for the formal experiment.

Break and eye tracker calibration. Following the practice session, participants can take a break and then the experimenter set up and calibrated the eye tracker for each participant to ensure data accuracy. Subsequently, the participants transitioned into the main experiment, applying the procedures and tasks they had become familiar with during the practice phase.

Formal experiment. In the formal experiment, subjects completed a 50-minute airspace monitoring of the main task and handled the unexpected warnings during the supervision. This duration was chosen to allow for sufficient data collection on DF while avoiding mental fatigue, which could introduce the confounding effects of fatigue-related increases in DF, making it harder to determine whether DF results from normal cognitive processes or mental exhaustion (Ferris et al., 2021). Detailed information about the tasks is shown below.

Airspace monitoring task. As shown in Figure 4, the experimental interface can be divided into three main areas, including the airspace monitoring area, warning display area, and warning-recording area. In practice, monitoring the radar screen and tracking the position, altitude, and speed of the aircraft are the basic tasks of ATCOs. Therefore, in this experiment, airspace monitoring is the subjects' main task. They need to supervise the aircraft in the airspace, recording the number of aircraft entering the airspace and the number of ordinary and heavy (H) aircraft successfully taking off and landing. Therefore, on the right side of the radar screen, there are five recorded contents, namely plane entering, successful take-off, successful take-off with H, successful landing, and successful landing with H.

Unexpected event handling. During the interface monitoring task, unexpected events were simulated through triggered warnings. In this experiment, receiving and recording warnings represented the handling of these events. If participants forgot the warning type, they could select the "others" button, indicating their attempt to comprehend the alert. To ensure exposure to various warning types and analyze participants' responses, a 50-minute air traffic control video was pre-recorded using the Endless ATC platform. It allows participants to perform a radar supervision task by monitoring the video. As depicted in the DF induction experiment phase shown in Figure 1, pre-experiment recording was conducted and critical aircraft control actions, such as altitude, speed, and heading, were designed to trigger warnings at random intervals, replicating unexpected real-world scenarios. All the eight types of warnings shown in Table A1 were induced in this video and shown in the recording area. Besides, only one warning appeared at a time, lasting for 10 seconds. The 10-second display time, supported by Friedman-Berg et al. (2008), who found that controllers need 10-12 seconds to recognize and respond to alerts, was also confirmed as sufficient in our practice phase for subjects to perceive and record the warnings. Each participant was exposed to 189 warnings over the 50-minute duration, reflecting the high frequency of alerts seen in real-world ATC environments (Ruskin et al., 2021) and ensuring sufficient data collection on detection failures, which are rare occurrences.

3.2.4. Warning frequency settings

Two types of warning frequency are set in the experiment, including continuous warnings and interval warnings, as shown in Figure 5. Continuous warnings can be defined as a subsequent warning that occurs immediately after a preceding warning, potentially overwhelming the operator's capacity to adequately respond to the primary task and unexpected warnings. By comparison, the subsequent warnings that appear at random intervals after the previous warning are regarded as interval warnings. Throughout the 50-minute experimental duration, participants were exposed to 189 warnings, comprising 68 continuous and 121 interval warnings. This continuous warning scenario is less intense compared to actual ATC situations where three or more consecutive warnings might occur.

3.3. Gaze data collection

In the experiment, detection failures were categorized based on subjects' gaze and response to warnings, as shown in Figure 6. OB was identified when there was no fixation on the warnings, indicating a complete miss. LBFTS occurred when subjects fixated on the warnings but failed to record them. MI errors were noted

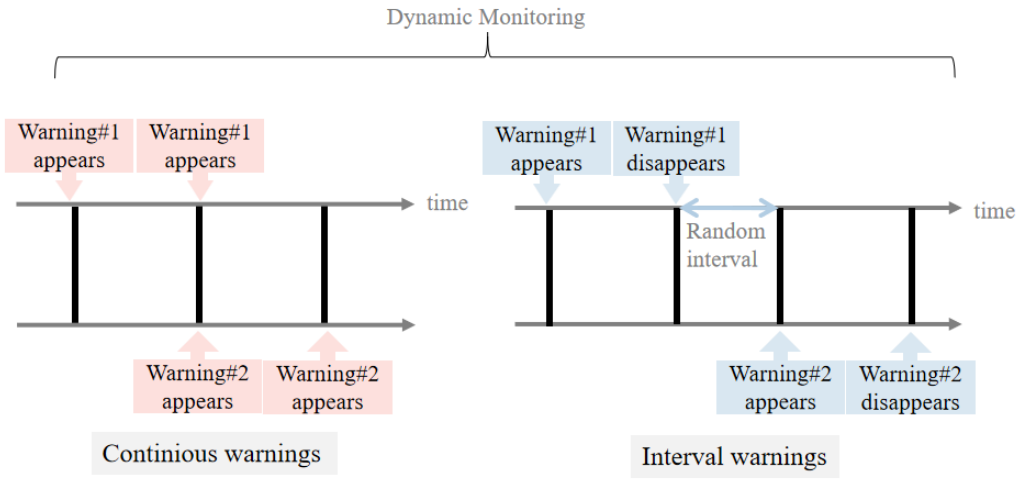


Figure 5: The settings of two types of warning frequency.

when subjects fixated on and recorded the warnings, but the recording was incorrect. Correction detection (CD) was achieved when subjects fixated on the warnings and recorded them accurately. Eye movement data were collected for each of the above four detection types. In cases of CD and MI, where subjects perceived and recorded the warnings, eye movements were tracked from the warning's appearance to its recording. For OB and LBFTS instances, where warnings were either perceived or responded to, eye movements were collected for the entire 10-second warning display duration.

After collecting the gaze data, the widely used eye-tracking features were extracted, including fixation count (FC) and fixation duration (FD) on the warning display area, the first view time (FVT) to the warnings, blink frequency (BF), mean saccade amplitude (MSA) and mean pupil diameter (MPD) in millimeters during the warning period. Besides, MPD includes mean left pupil diameter (MLPD) and mean right pupil diameter (MRPD).

After collecting the eye movement data, we applied the baseline normalization method to reduce individual differences in pupil diameter and blink frequency. Specifically, we consider the task-evoked pupillary response as the input of MPD. First, we set the baseline of MPD, which is in the thirty seconds after five minutes from the start of the experiment. Then, we compare the pupil diameter when the stimulus appears with that of the baseline, and the change between them is a task-evoked pupillary response. Similarly, we compare the BF when the stimulus appears with that of the baseline of BF, and the change between them is the task-evoked blink

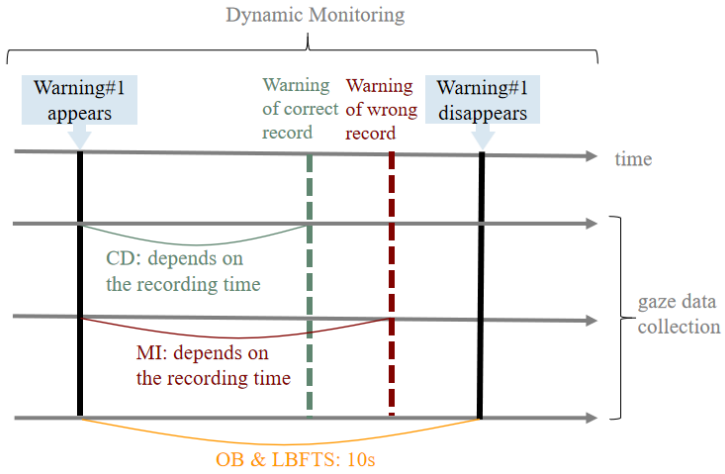


Figure 6: The eye movement data collection of different DF types during the experiment.

response.

3.4. Gaze dynamics across various DFs and warning frequencies

For statistical analysis, we consolidated the samples on a per-participant basis, producing a mean value for each individual before proceeding with significance testing. This approach is taken to ensure that the statistical significance is not artificially exaggerated by the larger number of samples. Meanwhile, samples from the same participant might be correlated with their unique behaviors and characteristics. Therefore, by averaging the data per participant, our results reflected true patterns rather than random variation within subjects and maintained the integrity of our statistical tests.

In our study, we compared the visual features under four detection scenarios (i.e. OB, LBFTS, MI, and CD). Initially, we employed the one-way Analysis of Variance (ANOVA) as our primary statistical method, which requires the data to be normally distributed as a prerequisite. However, the data did not meet the normality test, so we resorted to non-parametric methods, the Kruskal-Wallis H test. Additionally, we utilized the Mann-Whitney U test to evaluate the impact of warning frequency (i.e. continuous warnings and interval warnings) on gaze patterns for identical DF types. Specifically, the Kruskal-Wallis H test assesses median differences across three or more independent groups, while the Mann-Whitney U test compares medians between two independent samples.

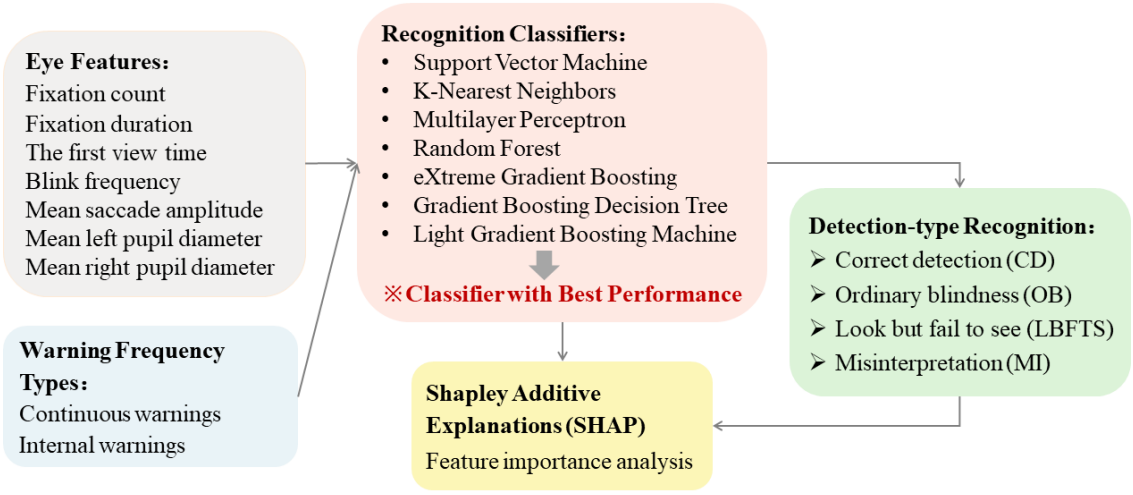


Figure 7: The process of DF-type recognition and analytics.

3.5. DF-type recognition and analytics

The process of DF-type recognition and analytics is shown in Figure 7. Two DF recognition tasks were compared: one with only eye movement inputs and another combining these with warning frequency types—'0' for interval and '1' for continuous warnings—providing insight into the impact of warning frequency on DF recognition. The specific features used in these recognition tasks are summarized in Table 1, detailing the relevant eye movement metrics and warning frequency information used as input for the models. The dataset consisted of 422 samples, including 108, 109, 95, 109 samples of OB, LBFTS, MI and CD, respectively, to ensure balanced sample sizes across all categories for fair comparison in the DF-recognition models. Meanwhile, 5-fold cross-validation was applied to train the recognition model, reducing overfitting by training and enhancing the reliability of our results. Seven machine learning methods, widely adopted in existing literature for their effectiveness, were employed and compared for DF-type recognition (Li et al., 2024; Shams et al., 2023). Three basic models like support vector machine (SVM), preferred for high-dimensional space handling; k-Nearest Neighbors (kNN), known for its simplicity and pattern recognition efficiency; and multilayer perceptron (MLP), a neural network proficient in learning non-linear relationships, were included (Cristianini & Shawe-Taylor, 2000). Additionally, we utilized advanced tree-based ensemble models such as random forest, eXtreme Gradient Boosting (XGBoost), and gradient boosting decision tree (GBDT), renowned since the early 2000s for their robustness in complex dataset analysis (Natras et al., 2022). Light Gradient Boosting Machine

(LightGBM), developed by Microsoft in 2017, stands out for its efficiency and performance in large dataset contexts (Ke et al., 2017). These models, selected for their classification prowess, were methodically assessed to identify the most fitting approach for DF-type recognition in our study. Furthermore, four evaluation metrics were used in this work, including accuracy, precision, recall, and F1 score (Yu et al., 2023). The output of each model is the classification of four detection types (i.e., OB, LBFTS, MI, and CD).

Table 1: The eye movement metrics and warning frequency feature used as input for machine learning models.

Features	Explanations
Fixation count (FC)	The number of fixations within the warning display area during a single warning display period (WDP).
Fixation duration (FD)	The total time spent fixating on the warning display area during a WDP in seconds.
Mean saccade amplitude (SA)	The average angular distance between consecutive fixations during a WDP.
Mean left/right pupil diameter (MLPD/MRPD)	The average size of the left or right pupil in millimeters during a WDP.
Blink frequency (BF)	The number of blinks during a WDP.
The first view time (FVT)	The time it takes to first fixate on the warning display area in seconds.
Warning frequency type (WFT)	The patterns of warning occurrence, either appearing immediately after the previous one (continuous warning) or following a time interval (interval warning).

The predictive outcomes of the recognition model with the best performance were interpreted using the Shapley Additive exPlanations (SHAP) method, which quantifies each feature's mean marginal contribution over all feature combinations (Zhou et al., 2021; Li et al., 2024). SHAP is a method derived from game theory to explain the output of machine learning models. It assigns each feature an importance value for a particular prediction, offering insights into how each feature contributes to the model's output. The core idea is based on the Shapley value, a concept from cooperative game theory that distributes payoffs fairly among players according to their contribution to the total payoff. The Shapley value for a feature value is defined as:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_S(x_{S \cup \{i\}}) - f_S(x_S)] \tag{1}$$

where: F is the set of all features, S is a subset of features excluding i, $|S|$ is the number of features in S, $|F|$ is the total number of features, $f_S(x_{S\cup\{i\}})$ is the model prediction with features in S plus i, $f_S(x_S)$ is the model prediction with features in S only. The overall prediction can be decomposed as the sum of the average model output and the SHAP values for each feature:

$$f(x) = \phi_0 + \sum_{i=1}^{|F|} \phi_i \tag{2}$$

where ϕ_0 is the average prediction of the model over the dataset. This approach aggregates these SHAP values, yielding a detailed measure of each feature's significance, clarifying their enhancing or diminishing effects on predictions. This process highlights the most influential features, and it also guides feature selection to enhance predictive accuracy.

4. Results

4.1. Gaze dynamics across various DFs and warning frequencies

4.1.1. Gaze patterns analysis across DF types

Following the collection of data from 26 participants, 109 instances of LBFTS, 328 occurrences of OB, and 95 MI were obtained. The data indicates that OB is the most common type of DF in response to warnings, appearing more than three times as often as LBFTS and MI. Furthermore, comprehensive Kruskal-Wallis tests of eye movement metrics across four warning detection types (i.e. OB, LBFTS, MI, CD) were conducted, as presented in Table 2. The statistical analysis, employing the Kruskal-Wallis test, identified significant variances in five metrics—FC, FD, MSA, MRLP, and FVT—across these detection types. The significance of these metrics reveals distinct visual behavior patterns linked to each detection type, verifying hypothesis 1 (H1). In contrast, BF and MLPD did not show statistically significant changes ($p = 0.445$ and $p = 0.088$, respectively), suggesting that these particular eye movement characteristics remain stable across varying detection scenarios. Further analysis through pairwise comparisons is also presented in Table 2 and allowed for an in-depth examination of each detection type's distinct eye-tracking features.

To further elaborate on the processed data and corresponding results, Figure 8 presents violin plots of five metrics with significant differences, as indicated in Table 2. These visualizations are closely tied to the data processing techniques described in Section 3.4, where samples were consolidated on a per-participant basis and averaged for each individual before undergoing statistical testing. The shapes of each violin

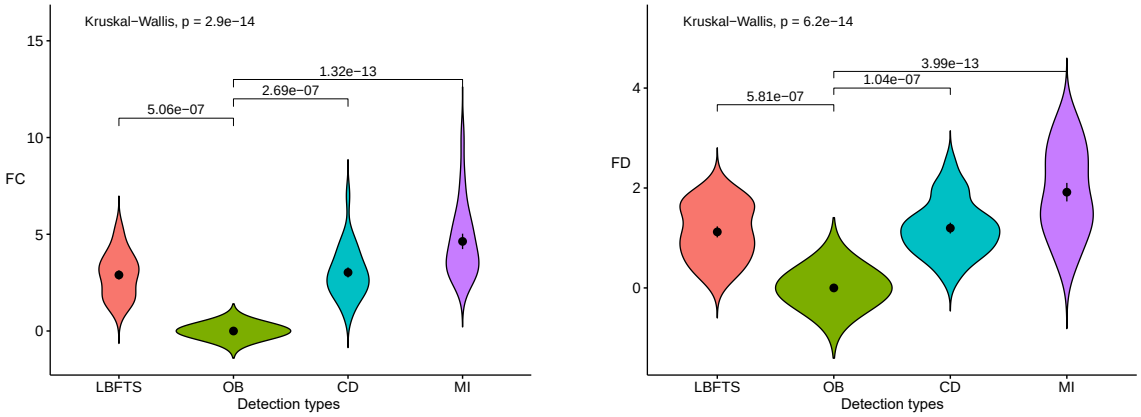
Table 2: Differential eye movement metrics across warning detection types.

Eye-tracking features	p-values in Kruskal-Wallis test (df=3)	Adjusted significance of pairwise comparisons in Kruskal-Wallis test					
	1-2-3-4	1-2	1-3	1-4	2-3	2-4	3-4
FC	0.000**	0.000**	0.000**	0.000**	0.077	1.000	0.085
FD	0.000**	0.000**	0.000**	0.000**	0.113	1.000	0.201
BF	0.445	-	-	-	-	-	-
MSA	0.000**	1.000	0.000**	0.000**	0.000**	0.000**	1.000
MLPD	0.088	-	-	-	-	-	-
MRPD	0.031*	1.000	1.000	0.563	0.306	0.015*	1.000
FVT	0.000**	0.000**	0.000**	0.000**	0.066	0.045*	1.000

1, 2, 3, and 4 denote OB, LBFTS, MI, and CD, respectively.
* p-value less than .05.
** p-value less than .01.

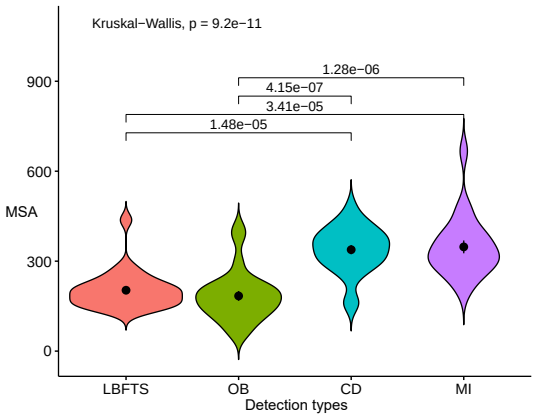
plot in Figure 8 reflect the distribution of specific features across all participants for each detection type, with wider areas indicating higher concentrations of participants sharing similar feature values, and narrower areas indicating fewer participants with those values. The p-values displayed at the top of each plot denote the overall statistical significance across the four detection types, determined using the Kruskal-Wallis test. Additionally, horizontal lines between detection types in each plot, along with their respective p-values, indicate the pairwise comparisons where significant differences were identified.

For both FC and FD metrics (see Figure 8a and Figure 8b), the OB detection type is markedly distinct, with significantly lower values indicating minimal or no fixations on warnings. This contrasts with the LBFTS and CD conditions, which display comparable distributions, suggesting a moderate level of attention to the warnings. The MI condition stands out with its broader distribution, skewing towards higher fixation counts and longer durations, indicative of increased attention and engagement with the warnings. The significant differences between the OB condition and the other three detection types are underscored by the lines with numerical p-values, confirming these observations' statistical significance according to the Kruskal-Wallis test. For the MSA feature (see Figure 8c), the distribution across the detection types varies, with OB having a lower range compared to others. The plot shows that MSA is significantly higher for CD and MI compared to LBFTS and OB, as denoted by the lines connecting these conditions and the p-values above them, indicating that saccades are larger when warnings are cognitively processed. Additionally, as shown in Figure 8d, the MRPD plot illustrates varying distributions of pupil dilation across detection types, with LBFTS and CD showing a statistically

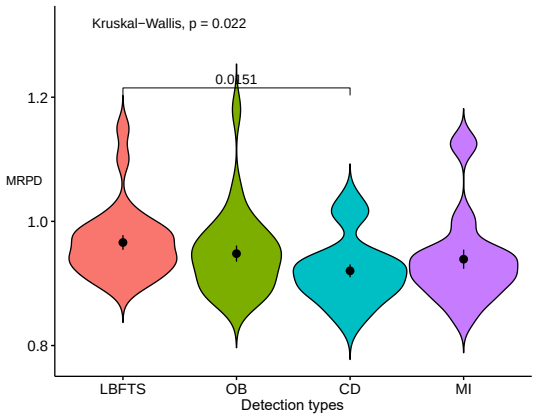


(a) FC

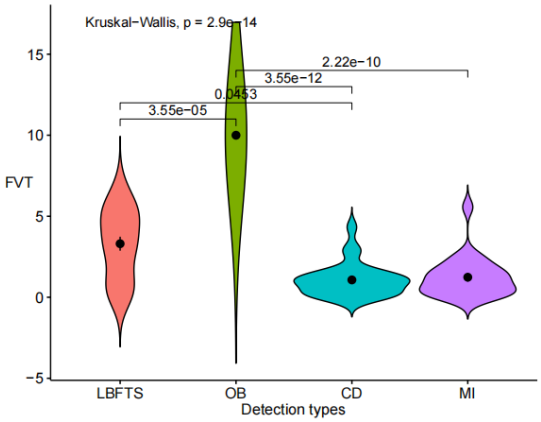
(b) FD



(c) MSA



(d) MRPD



(e) FVT

Figure 8: The violin plots of five metrics with significant differences across the detection types.

significant difference, as denoted by the p-value of 0.015 highlighting the divergence. This disparity suggests that the cognitive demands of conflict detection in the CD condition may be higher than in the LBFTS condition. Consequently, the variation in MRPD between these two types of detection could serve as a discriminative feature, potentially useful for differentiating between the cognitive processes associated with LBFTS and CD events. In our analysis of the FVT feature, as illustrated in Figure 8e, the OB condition stands out significantly as it completely missed the warning within the 10-second timeframe after onset, showing no reaction to the alert. However, interestingly, the first view time for LBFTS is distributed more variably and shows a substantial difference from both CD and MI. The divergence between LBFTS and CD is statistically significant, with a p-value of 0.045. In comparison, CD and MI display relatively quicker and more concentrated first view times, indicating a more prompt and focused response to the warnings, whereas the first view times for LBFTS are more spread out, implying a less consistent response pattern.

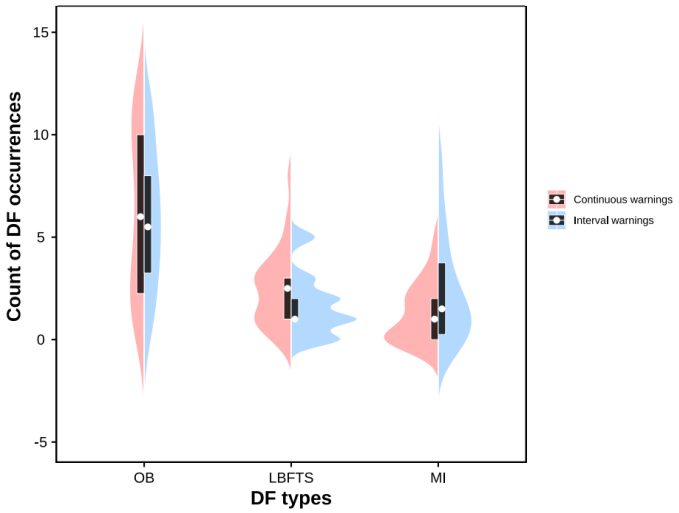


Figure 9: The split violin plot for comparative distribution of DF types across warning frequency conditions.

4.1.2. Effects of warning frequency on DFs reflected by eye metrics

In our study of how warning frequencies influence DF occurrences through eye movement effects, we initially analyzed the distribution of DF types across two warning frequencies, as shown in Figure 9. The figure displays the frequency of each DF type among 26 subjects under continuous and interval warning scenarios. Black bars represent data spread and median values. Notably, the occurrence of OB exhibits a

Table 3: p-values of Mann-Whitney U test for continuous and interval warnings across three DF types and various eye-tracking features.

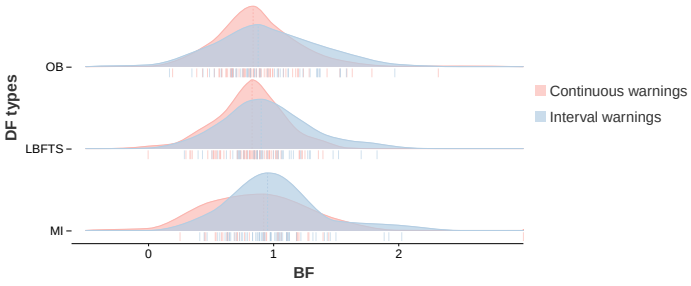
Eye-tracking features	p-values of Mann-Whitney U test for continuous and interval warnings across three DF types		
	OB	LBFTS	MI
FC	1.000	0.462	0.415
FD	1.000	0.980	0.124
BF	0.046*	0.038*	0.312
MSA	0.744	0.037*	0.048*
MLPD	0.398	0.314	0.487
MRPD	0.402	0.436	0.863
RT	1.000	0.643	0.010**

* p-value less than .05.
** p-value less than .01.

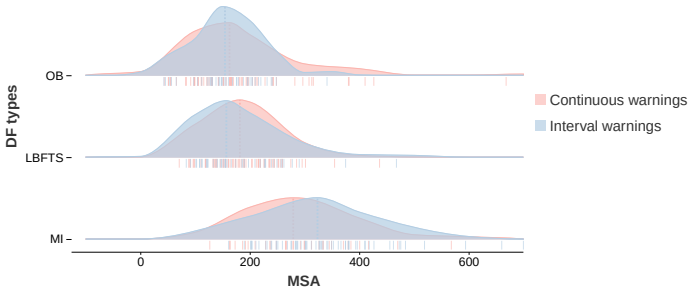
dispersed distribution in both warning types, indicating diverse operator experiences with OB. Conversely, LBFTS and MI demonstrate more concentrated distributions, generally centering around lower occurrences. Meanwhile, OB and LBFTS, especially LBFTS, arise more frequently under continuous warnings, suggesting heightened susceptibility to such conditions. Conversely, MI occurrences are less frequent during continuous warnings, indicating a relative insensitivity to continuous warning presence.

Furthermore, the Mann-Whitney U test was employed to assess the influence of warning frequency (continuous versus interval warnings) on gaze patterns for identical DF types. The results are detailed in Table 3, where significant differences were observed. For example, a p-value of 0.046 indicates a notable divergence in blink frequency (BF) between continuous and interval warnings under the OB condition. Metrics such as BF, MSA, and FVT exhibited statistical significance $p < 0.05$, demonstrating their ability to differentiate between warning frequencies across the detection types tested. This provides strong evidence in support of Hypothesis 2 (H2). Figure 10 complements these findings by displaying ridgeline plots of the three significant gaze metrics, where density lines in each plot capture the overall sample distribution for each warning frequency type, and the median line marks the median gaze data.

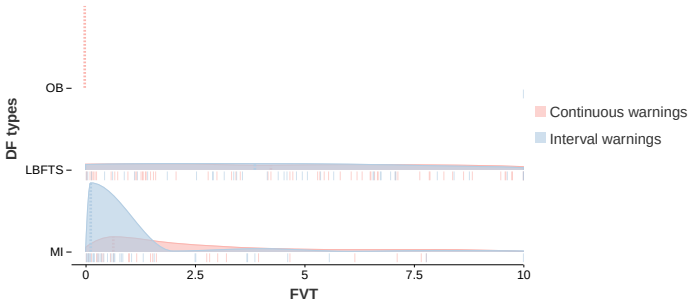
Figure 10a presents a ridgeline plot of BF differences among DF types (OB, LBFTS, and MI). The median lines, when evaluated alongside Table 1, reveal a distinct BF distribution for OB and LBFTS under varying warning frequencies. Under continuous warnings, subjects show reduced and more uniform BF, implying poten-



(a) Blink frequency



(b) Mean saccade amplitude



(c) The first view time

Figure 10: The ridgeline plots of three metrics with significant differences between continuous and interval warnings under different DF types.

tial warning fatigue and a consequent decrease in blinking as subjects strive to remain vigilant. Conversely, interval warnings exhibit a varied BF distribution, suggesting a more variable blinking pattern. For MI, continuous warnings slightly reduce BF but not significantly when compared to interval warnings, suggesting minimal impact on blink rate in this context.

Figure 10b illustrates that, in LBFTS scenarios, the MSA is significantly larger during continuous warnings than interval warnings, possibly reflecting a cognitive response to frequent alerts that broadens the visual search to assimilate information overload. Conversely, MI scenarios show a reduced MSA with continuous warnings, potentially a cognitive strategy for a focused attentional state to constrain attention to critical elements amidst frequent stimuli. These contrasting patterns in saccadic responses could reflect adaptive strategies to manage cognitive load under different warning frequencies.

As depicted in Figure 10c, the OB scenario shows no FVT distribution due to the total neglect of warnings. In LBFTS scenarios, the widespread distribution and larger median of FVT indicate a consistent delay in noticing warnings, irrespective of warning frequency. This points towards a cognitive processing challenge where subjects notice the warnings but fail to recognize their significance promptly, leading to delayed or inadequate processing of the visual warnings. Conversely, in MI scenarios, FVT is markedly higher under continuous warnings than interval warnings. The slowing initial response reflects deeper cognitive engagement and careful assessment of each warning.

4.2. DF-type recognition and analytics

4.2.1. Gaze-based DF-type recognition

Table 4 presents a comparative analysis of seven machine learning models for categorizing eye-tracking data into four detection types, highlighting GBDT as the most effective model with uniform scores of 0.74 across accuracy, precision, recall, and F1 score. This indicates its superior capability in the four-type classification. RF and lightGBM also perform well, particularly RF, which matches GBDT in accuracy, precision, and recall. Among basic models, MLP outperforms SVM and kNN. This comparison illustrates the advanced models' proficiency in handling complex multi-class eye-tracking data classifications.

Further, as detailed in Session 4.1.2, warning frequencies influence DF occurrences, prompting us to integrate warning frequency features into DF type identification to assess potential improvements in results. The results are shown in Table 5, indicates an enhancement in classification accuracy when including the warning frequency feature. The RF model stands out with the highest scores of 0.80 for

Table 4: Performance of the compared models for classifying eye-tracking features into four detection types.

	SVM	kNN	MLP	RF	XGBoost	GBDT	LightGBM
Accuracy	0.67	0.65	0.69	**0.74**	0.67	**0.74**	0.73
Precision	0.68	0.65	0.69	**0.74**	0.66	**0.74**	0.73
Recall	0.67	0.66	0.69	**0.74**	0.67	**0.74**	0.73
F1 score	0.66	0.64	0.67	0.73	0.66	**0.74**	0.73

Table 5: Performance of the compared models for classifying eye-tracking features and warning frequency types into four detection types.

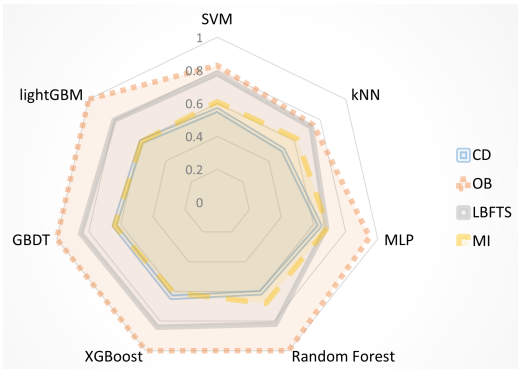
	SVM	kNN	MLP	RF	XGBoost	GBDT	LightGBM
Accuracy	0.69	0.63	0.68	**0.79**	0.76	0.77	0.74
Precision	0.69	0.66	0.67	**0.80**	0.77	0.78	0.75
Recall	0.69	0.64	0.68	**0.79**	0.77	0.78	0.74
F1 score	0.69	0.60	0.66	**0.79**	0.77	0.78	0.74

precision and 0.79 for accuracy, recall, and F1 score, underscoring the benefit of integrating warning frequency context. These results verified our hypothesis 3 (H3), the gaze features and warning frequency can be used to recognize detection types.

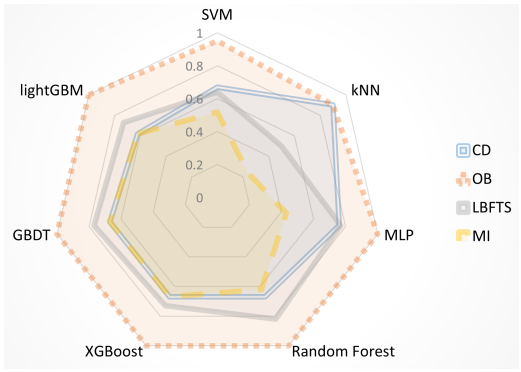
After evaluating overall results, a model performance comparison across the detection categories–CD, OB, LBFTS, and MI is conducted, corresponding to the results of Table 5, as shown in Figure 11. Figure 11 shows notable variations in model effectiveness across these categories, with each model demonstrating varying degrees of proficiency for the four detection types. OB is identified as the category with the highest prediction accuracy, succeeded by LBFTS, CD, and MI in that order. This differentiation indicates that OB and LBFTS possess distinct eye movement characteristics, resulting in markedly higher prediction accuracy. Subsequently, a detailed analysis of the prediction results for each detection type is provided as follows.

Within the OB type, defined by the lack of fixation on warnings, the models, particularly the four advanced tree-based models, exhibit outstanding accuracy, achieving precision and recall scores of up to 1.00. This exceptional performance highlights the unique gaze characteristics associated with OB, such as absent fixations and extended first view time as clarified by prior research (Hergovich & Oberfichtner, 2016), providing distinct indicators for classification.

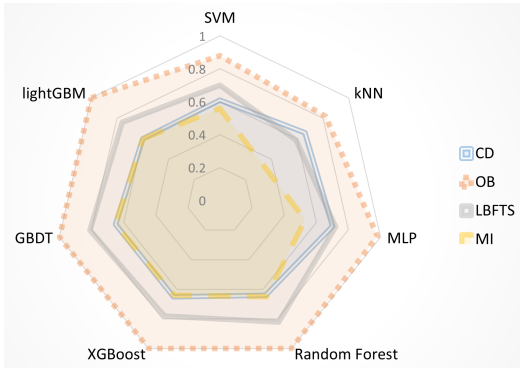
LBFTS exhibits unique gaze patterns, differentiating it from CD and MI, all of which involve fixations on warnings. The GBDT and RF models excel in LBFTS detection: GBDT scores 0.82 in precision, recall, and F1, whereas RF shows a pre-



(a) Precision



(b) Recall

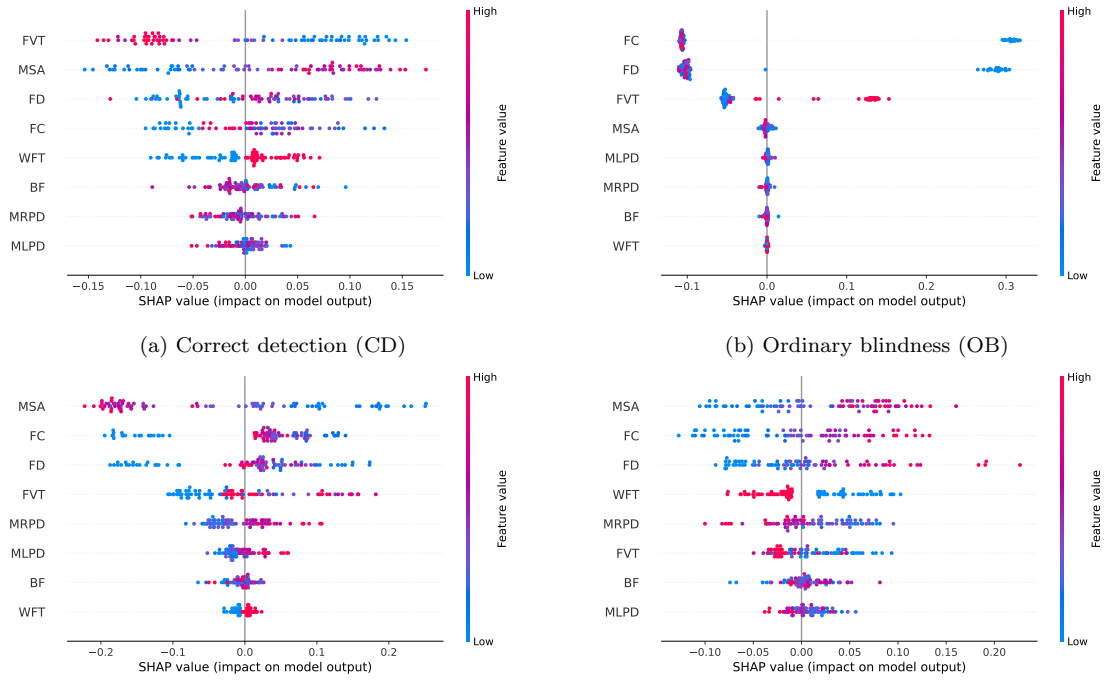


(c) F1 score

Figure 11: A comparative analysis of model performances from a single specific detection category.

cision of 0.85, recall of 0.77, and F1 of 0.81. These outcomes emphasize LBFTS's distinct eye movements compared to CD's focused attention and MI's attentional errors. As supported by findings in Session 4.1.1, LBFTS correlates with reduced saccade amplitude, enlarged pupil diameter, and prolonged first view time. These findings underscore the criticality of eye movement patterns in identifying LBFTS, affirming their integral role in DF classification.

The MI and CD categories, representing incorrect and correct warning recordings, respectively, exhibit distinct predictive challenges. MI slightly surpasses in precision, with MLP scoring 0.69, compared to CD's highest of 0.64. In contrast, CD excels in the recall, with kNN achieving 0.9, outdoing MI's best of 0.67 by XGBoost. Assessing F1 scores, CD models have a slight advantage, with MLP for CD reaching 0.7 versus MI's top score of 0.65 by GBDT. These results suggest CD type is more distinguishable in eye movement analysis, yet the classification outcomes for MI also remain significantly high.



(a) Correct detection (CD)

(b) Ordinary blindness (OB)

(c) Look but fail to see (LBFTS) error

(d) Misinterpretation (MI)

Figure 12: Beeswarm plot of importance ranking of the eye-tracking and warning frequency type (WFT) features.

4.2.2. Feature importance analysis by SHAP

SHAP is used to interpret and explain the feature importance and decision-making process within a random forest model, selected for its optimal recognition performance. Figure 12 presents SHAP value distributions, denoting the importance of a range of input features, including eye movement metrics and warning frequency types (WFT), on the four-category classification. The SHAP values are each sub-figure (a-d) corresponds to one detection type, illustrating the impact of various features on model output. The color spectrum from blue to red depicts feature values ranging from low to high. SHAP values on the horizontal axis quantify the influence of each feature on the model's prediction, with positive values indicating an increase and negative values a decrease in the likelihood of the target class. Features are ranked vertically by their importance, with the most impactful at the top.

A comprehensive analysis of these four sub-figures highlights key features—FC, FVT, and MSA—as effective in identifying four detection failure (DF) types, followed by FD and WFT. Specifically, in subfigure 12b, the defining characteristic of

OB—the absence of fixations and first view times on warnings—makes it easily recognizable. This is especially reflected in the tightly mixed red and blue points for FC and FD. Since OB is characterized by zero fixations on warnings, any non-zero value of FC and FD, regardless of whether it is small or large, contributes similarly to the classification of instances as non-OB. Then, in subfigure 12a, CD's quicker response time is captured by FVT as a critical feature, which is consistent with the findings in Session 4.1.1. Moreover, for LBFTS and MI, as shown in subfigure 12c and 12d, the scanning range is a differentiator. Despite both perceiving warnings, LBFTS shows smaller saccade amplitude and fewer fixations, making MSA, FC, and FD useful in distinguishing it from MI. Additionally, compared to the subfigures of 12a and 12d, two warning frequency types affect the recognitions of CD and MI, and this impact is interesting. Continuous warnings, contrary to expectations, reduce MI occurrences. This phenomenon suggests that a constant flow of alerts keeps operators vigilantly responsive. In contrast, interval warnings disrupt operators' expectation patterns of operators, potentially leading to inconsistent cognitive engagement and misinterpretations. This is consistent with the top-down stimulus-related expectations elaborated in Hutchinson (2019).

5. Discussions

5.1. Gaze dynamics and importance analytics in DF-type recognition

The statistical results in Section 4.1.1 reveal that four detection types exhibit significantly different gaze patterns, demonstrating that non-perception, unaware perception, and aware perception of warnings are characterized by unique gaze dynamics. Consequently, each DF type possesses distinct gaze indicators, yielding robust recognition outcomes. Notably, OB achieves the highest recognition accuracy, succeeded by LBFTS, CD, and MI. The subsequent discussion elaborates on the specific eye movement characteristics derived from the statistical analyses for each category, which align with the feature importance analysis provided by SHAP.

The occurrence of OB can be characterized by statistically significant eye-tracking features, including the absence of fixations, first view time to warnings, and notably reduced saccade amplitude, as indicated by the statistical analysis in Section 4.1.1. These metrics collectively help to identify OB and summarize the lack of visual and cognitive engagement with critical warnings, consistent with prior research (Hergovich & Oberfichtner, 2016). No response reflects a failure to perceive the warnings, suggesting the warning display needs to be improved or the user's attention resources are insufficient (Richards et al., 2012). The reduced saccade amplitude signifies a constrained visual search, implying that the user's gaze remains fixated within a

limited area, potentially missing the peripheral warnings (Li et al., 2023). Extended first view time further affirms the non-recognition of alerts, indicating a delayed or non-existent cognitive processing of the warnings. These features collectively provide a composite signal of OB, underscoring the need for improved warning strategies to capture and maintain operator attention.

LBFTS can be differentiated from MI and correct detection through distinct eye-tracking metrics identified in the statistical analysis: a reduced saccade amplitude distinguishes LBFTS from MI, while delayed first view time and an increased pupillary diameter separate LBFTS from correct detection. The smaller saccade magnitude in LBFTS compared to MI suggests a more focused yet potentially insufficient visual search area, increasing the likelihood of overlooking critical stimuli (Bodala et al., 2016). As highlighted in Moacdieh et al. (2023), the increased pupillary diameter indicates a heightened cognitive workload. When the workload continually increases, it can lead to cognitive fatigue and resource misallocation, ultimately reducing the efficiency of information processing, as supported by research on workload and fatigue impacts on performance (Fan & Smith, 2017). Therefore, the increased workload may coincide with insufficient information processing, particularly when the operator is under strain, leading to the LBFTS error. This differentiation between LBFTS and correct detection becomes apparent when increased pupil diameter is correlated with insufficient processing capacity. Furthermore, the slower first view time during LBFTS occurrences reflects a lag in cognitive recognition of the warnings (Renata et al., 2018), further underscoring the impact of cognitive overload on performance. These metrics imply the attentional and cognitive state during LBFTS errors and could guide the development of targeted strategies to deal with the task load of controllers.

Records can indicate MI versus correct detection in certain contexts: incorrect records imply warning misinterpretation, while accurate ones suggest proper detection and understanding, as noted by Hollnagel (2000). While the statistical analysis in Section 4.1.1 did not reveal significant differences in eye-tracking metrics between MI and CD, the SHAP analysis provides valuable insights into the distinct contributions of key eye-tracking features for their classification. For CD, the key indicator is the first view time, which suggests a rapid and efficient response to warnings. This aligns with the necessity for prompt recognition and reaction in situations requiring accurate interpretation (Wilkinson & Seales, 1978). On the other hand, MI is primarily distinguished by the scanning range, indicating a potentially less focused visual search. According to Navarro-Cebrian et al. (2013), uncertain responses in MI cases were often due to failures at the time of stimulus processing, marked by lower amplitude sensory event-related potentials. This scanning behavior in MI

could be indicative of uncertainty or confusion, leading to erroneous interpretations of warning signals. These variations in eye movement patterns, revealed by advanced modeling techniques, not only enhance our understanding of detection behaviors but also indicate potential areas for targeted training and system improvements in ATC operations.

5.2. Effects of warnings frequency on DF types

Among the three DF types, OB and LBFTS occurrences differ between continuous and interval warnings, particularly with LBFTS being more common during continuous warnings. It indicates that continuous warnings affect operators' awareness of the warning, resulting in more unaware perceptions of warnings. Moreover, eye-tracking metrics further clarify the effects of warning frequency on DF types. Based on the statistical findings from Section 4.1.2, the significant reduction in blink frequency under continuous warnings for OB and LBFTS suggests the occurrence of cognitive tunneling. The concept of cognitive tunneling is well-documented in high workload environments, where individuals, overwhelmed by simultaneous stimuli, narrow their focus on information from specific areas of a display, often ignoring broader contextual stimuli (Thomas & Wickens, 2001). This attentional narrowing explains both OB and LBFTS: under continuous warnings, operators may either miss warnings entirely (OB) or fail to process them fully despite seeing them (LBFTS), as their cognitive resources become increasingly focused on previously processed tasks or stimuli, leaving poorer information extraction and situation awareness to address new warnings effectively. The reduction in blink frequency during successive warnings reflects this narrowing of attention and provides valuable insights into the operator's attentional state and ability to manage multiple stimuli. Additionally, for LBFTS, the increased saccade amplitude under continuous warnings indicates a broader visual search, potentially as a compensatory response to processing excessive information (Li et al., 2023).

In recognizing four DF types, warning frequency mainly impacts MI and CD recognition. Interestingly, the presence of continuous warnings diminishes occurrences of MI. It suggests that a consistent flow of warnings enhances the operators' foresight of warning implications in a state of heightened vigilance. Observations in Session 4.1.2 indicate that under continuous warnings, MI is associated with reduced saccade amplitude, pointing to a more focused attentional state. In contrast, interval warnings may cause attentional lapses and unstable cognitive engagement, increasing misinterpretations. This aligns with top-down stimulus-related expectations, where attention and prior knowledge shape perception and cognitive processing (Hutchinson, 2019). In other words, operators' improved response to continuous warnings is

shaped by their cognitive expectations and experience, reducing misinterpretations by focusing attention and using prior knowledge to enhance vigilance and perception accuracy. In summary, continuous warnings reduce operators' alert awareness but improve their implications foresight. Thus, warning frequency should be adjusted based on the targeted DF type for optimal mitigation.

6. Conclusions

This study aims to systematically classify DF in response to ATC warnings, developing a four-phase framework for identifying effective DF-type indicators and machine-learning models. The framework includes gaze-behavioral DF classification, DF induction experiment, gaze dynamics across various DFs and warning frequencies, and DF-type recognition and analytics. Additionally, it examines the influence of warning frequency on DFs, as evident in eye-tracking data. The study identifies distinct gaze dynamics between non-perception, unaware perception, and aware perception, with the gradient-boosting decision tree model demonstrating a notable 74% precision in four-category recognition. A further enhancement is observed when warning frequency is considered, with the random forest model achieving an 80% precision. This study offers valuable theoretical and practical insights into advancing DF recognition and intervention in complex environments.

At the theoretical level, this study extends the application of SA theory in complex systems, particularly through the systematic classification of DF types and the provision of specific eye-tracking indicators and attributional explanations for each type. This research provides new insights into the interaction between situational factors, particularly warning frequency and gaze behavior, and their influence on visual perception and awareness. It highlights that fixation patterns, saccade amplitude, and pupil diameter collectively reveal how operators balance perception, cognitive engagement, and stimuli expectations, thereby affecting their processing of external stimuli. Furthermore, a consistent flow of warnings weakens operators' awareness of warnings yet enhances the foresight of warning implications. These insights offer valuable theoretical guidance for the evaluation of warning detection in complex environments, enabling better anticipation and response to fluctuations in operator detection behaviors.

At a practical level, this study provides critical insights for developing dynamic, adaptive warning systems encompassing impact mechanisms, DF recognition, and intervention strategies. The research reveals the complex relationship between warning frequency and operator situational awareness, proposing that warning systems should dynamically adjust frequency mechanisms based on real-time cognitive states. By

monitoring gaze dynamics, the system can accurately identify different types of DF and implement customized intervention strategies, such as increasing task prompts or adjusting the complexity of warnings. This integrated approach enhances the responsiveness of warning systems and offers practical solutions for real-time monitoring and human-computer collaboration in complex environments.

There are certain limitations that future research should address. The current experimental setup focuses on basic ATC tasks, suggesting future research explores more complex and specific ATC scenarios to understand the broader spectrum of DF phenomenon. Meanwhile, the challenges of data collection in critical ATC scenarios and the training required for subjects led to a small sample size. Additionally, exploring a broader range of warning frequencies, like variations in continuous warning intensity will be crucial for developing comprehensive strategies to mitigate DFs and enhance safety and efficiency in ATC operations.

Acknowledgments

This work was supported by the Hong Kong Polytechnic University under Grant P0038827 and Grant P0038933. This study has been granted human ethics approval from the PolyU Institutional Review Board of The Hong Kong Polytechnic University (IRB Reference Number: HSEARS20211117002).

Data Availability Statement

The datasets and source code used in this study are publicly available at [https://github.com/Jasmine-LZM/Eye-Tracking-Data-and-Source-Code-for-DF-Recognition]. This repository ensures transparency and allows the research community to verify and build upon our work.

References

Akhawe, D., & Felt, A. P. (2013). Alice in warningland: a {Large-Scale} field study of browser security warning effectiveness. In *22nd USENIX security symposium (USENIX Security 13)* (pp. 257–272).

Anderson, B. B., Jenkins, J. L., Vance, A., Kirwan, C. B., & Eargle, D. (2016). Your memory is working against you: How eye tracking and memory explain habituation to security warnings. *Decision Support Systems*, *92*, 3–13.

Ball, F., Elzemann, A., & Busch, N. A. (2014). The scene and the unseen: Manipulating photographs for experiments on change blindness and scene memory: Image manipulation for change blindness. *Behavior research methods*, *46*, 689–701.

Bodala, I. P., Abbasi, N. I., Sun, Y., Bezerianos, A., Al-Nashash, H., & Thakor, N. V. (2017). Measuring vigilance decrement using computer vision assisted eye tracking in dynamic naturalistic environments. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 2478–2481). IEEE.

Bodala, I. P., Li, J., Thakor, N. V., & Al-Nashash, H. (2016). Eeg and eye tracking demonstrate vigilance enhancement with challenge integration. *Frontiers in human neuroscience*, *10*, 273.

Brinton Anderson, B., Vance, A., Kirwan, C. B., Eargle, D., & Jenkins, J. L. (2016). How users perceive and respond to security messages: a neurois research agenda and empirical study. *European Journal of Information Systems*, *25*, 364–390.

Bruder, C., & Hasse, C. (2019). Differences between experts and novices in the monitoring of automated systems. *International Journal of Industrial Ergonomics*, *72*, 1–11.

Bruder, C., & Hasse, C. (2020). What the eyes reveal: Investigating the detection of automation failures. *Applied ergonomics*, *82*, 102967.

Castritius, S.-M., Schubert, P., Dietz, C., Hecht, H., Huestegge, L., Liebherr, M., & Haas, C. T. (2021). Driver situation awareness and perceived sleepiness during truck platoon driving–insights from eye-tracking data. *International Journal of Human–Computer Interaction*, *37*, 1467–1477.

Corver, S. C., & Aneziris, O. N. (2015). The impact of controller support tools in enroute air traffic control on cognitive error modes: A comparative analysis in two operational environments. *Safety science*, *71*, 2–15.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human factors*, *49*, 564–572.

Eißfeldt, H., Grasshoff, D., Hasse, C., Hörmann, H.-J., Schulze Kissing, D., Stern, C., Wenzel, J., & Zierke, O. (2010). Aviator 2030-ability requirements in future atm systems ii: Simulations and experiments, .

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human factors*, *37*, 32–64.

Fan, J., & Smith, A. P. (2017). The impact of workload and fatigue on performance. In *Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers 1* (pp. 90–105). Springer.

Ferris, J. R., Tomlinson, M. A., Ward, T. N., Pepin, M. E., & Malek, M. H. (2021). Reduced electromyographic fatigue threshold after performing a cognitive fatiguing task. *The Journal of Strength & Conditioning Research*, *35*, 267–274.

Friedman-Berg, F., Allendoerfer, K., & Pai, S. (2008). Nuisance alerts in operational atc environments: Classification and frequencies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 104–108). SAGE Publications Sage CA: Los Angeles, CA volume 52.

Friedrich, M., Rußwinkel, N., & Möhlenbrink, C. (2017). A guideline for integrating dynamic areas of interests in existing set-up for capturing eye movement: Looking at moving aircraft. *Behavior research methods*, *49*, 822–834.

Goldberg, J. H., & Kotval, X. P. (1998). Eye movement-based evaluation of the computer interface. *Advances in occupational ergonomics and safety*, (pp. 529–532).

Han, S., Li, F., Lee, C.-H., Wang, T., & Diaconeasa, M. A. (2024). Mirror the mind of crew: Maritime risk analysis with explicit cognitive processes in a human digital twin. *Advanced Engineering Informatics*, *62*, 102746.

Hergovich, A., & Oberfichtner, B. (2016). Magic and misdirection: The influence of social cues on the allocation of visual attention while watching a cups-and-balls routine. *Frontiers in Psychology*, *7*, 761.

Hollnagel, E. (2000). Looking for errors of omission and commission or the hunting of the snark revisited. *Reliability Engineering & System Safety*, *68*, 135–145.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. oup Oxford.

Hou, G., Dong, Q., & Wang, H. (2024). The effect of dynamic effects and color transparency of ar-hud navigation graphics on driving behavior regarding inattentional blindness. *International Journal of Human–Computer Interaction*, (pp. 1–12).

Hutchinson, B. T. (2019). Toward a theory of consciousness: a review of the neural correlates of inattentional blindness. *Neuroscience & Biobehavioral Reviews*, *104*, 87–99.

Jones, K. S., Lodinger, N. R., Widlus, B. P., Namin, A. S., & Hewett, R. (2021). Do warning message design recommendations address why non-experts do not protect themselves from cybersecurity threats? a review. *International Journal of Human–Computer Interaction*, *37*, 1709–1719.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Kontogiannis, T., & Malakis, S. (2009). A proactive approach to human error detection and identification in aviation and air traffic control. *Safety Science*, *47*, 693–706.

Li, F., Chen, C.-H., Lee, C.-H., & Feng, S. (2022). Artificial intelligence-enabled non-intrusive vigilance assessment approach to reducing traffic controller's human errors. *Knowledge-Based Systems*, *239*, 108047.

Li, F., Chen, C.-H., Xu, G., & Khoo, L.-P. (2020). Hierarchical eye-tracking data analytics for human fatigue detection at a traffic control center. *IEEE Transactions on Human-Machine Systems*, *50*, 465–474.

Li, Z., Li, R., Yuan, L., Cui, J., & Li, F. (2024). A benchmarking framework for eye-tracking-based vigilance prediction of vessel traffic controllers. *Engineering Applications of Artificial Intelligence*, *129*, 107660.

Li, Z., Li, Z., & Li, F. (2023). Visual attention analytics for individual perception differences and task load-induced inattentional blindness. In *International Conference on Human-Computer Interaction* (pp. 71–83). Springer.

Liao, C.-W., & Chiang, T.-L. (2016). Reducing occupational injuries attributed to inattentional blindness in the construction industry. *Safety science*, *89*, 129–137.

Liu, Y., Trapsilawati, F., Lan, Z., Sourina, O., Johan, H., Li, F., Chen, C.-H., & Mueller-Wittig, W. (2020). Human factors evaluation of atc operational procedures in relation to use of 3d display. In *Advances in Human Factors of Transportation: Proceedings of the AHFE 2019 International Conference on Human Factors in Transportation, July 24-28, 2019, Washington DC, USA 10* (pp. 715–726). Springer.

Lyu, M., Li, F., Lee, C.-H., & Chen, C.-H. (2024a). Valio: Visual attention-based linear temporal logic method for explainable out-of-the-loop identification. *Knowledge-Based Systems*, (p. 112086).

Lyu, M., Li, F., Qu, X., & Li, Q. (2024b). Flashlight model: integrating attention distribution and attention resources for pilots' visual behaviour analysis and performance prediction. *International Journal of Industrial Ergonomics*, *103*, 103630.

Ma, S., Zhang, W., Yang, Z., Kang, C., Wu, C., Chai, C., Shi, J., Zeng, Y., & Li, H. (2021). Take over gradually in conditional automated driving: the effect of two-stage warning systems on situation awareness, driving stress, takeover performance, and acceptance. *International Journal of Human–Computer Interaction*, *37*, 352–362.

Mack, A., & Rock, I. (1998). Inattentional blindness: Perception without attention. *Visual attention*, *8*.

Mengtao, L., Fan, L., Gangyan, X., & Su, H. (2023). Leveraging eye-tracking technologies to promote aviation safety-a review of key aspects, challenges, and future perspectives. *Safety science*, *168*, 106295.

Moacdieh, N. M., Dibo, M., Halabi, Z., & Antoun, J. (2023). Eye tracking to evaluate the effectiveness of electronic medical record training. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications* (pp. 1–7).

Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble machine learning of random forest, adaboost and xgboost for vertical total electron content forecasting. *Remote Sensing*, *14*, 3547.

Navarro, J., Deniel, J., Yousfi, E., Jallais, C., Bueno, M., & Fort, A. (2019). Does false and missed lane departure warnings impact driving performances differently? *International Journal of Human–Computer Interaction*, *35*, 1292–1302.

Navarro-Cebrian, A., Knight, R. T., & Kayser, A. S. (2013). Error-monitoring and post-error compensations: dissociation between perceptual failures and motor errors with and without awareness. *Journal of Neuroscience*, *33*, 12375–12383.

Nikkei (2024). Haneda air traffic control missed warning alert in jal collision. https://asia.nikkei.com/Spotlight/Japan-plane-crash/Haneda-air-traffic-control-missed-warning-alert-in-JAL-collision.

Orquin, J. L., & Loose, S. M. (2013). Attention and choice: A review on eye movements in decision making. *Acta psychologica*, *144*, 190–206.

Papastavrou, J. D., & Lehto, M. R. (1995). A distributed signal detection theory model: Implications for the design of warnings. *International Journal of Occupational Safety and Ergonomics*, *1*, 215–234.

Papastavrou, J. D., & Lehto, M. R. (1996). Improving the effectiveness of warnings by increasing the appropriateness of their information content: some hypotheses about human compliance. *Safety science*, *21*, 175–189.

Park, S., Park, C. Y., Lee, C., Han, S. H., Yun, S., & Lee, D.-E. (2022). Exploring inattentional blindness in failure of safety risk perception: Focusing on safety knowledge in construction industry. *Safety science*, *145*, 105518.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*, 372.

Renata, V., Li, F., Lee, C.-H., & Chen, C.-H. (2018). Investigation on the correlation between eye movement and reaction time under mental fatigue influence. In *2018 International Conference on Cyberworlds (CW)* (pp. 207–213). IEEE.

Richards, A., Hannon, E. M., & Vitkovitch, M. (2012). Distracted by distractors: Eye movements in a dynamic inattentional blindness task. *Consciousness and cognition*, *21*, 170–176.

Ruscio, D., Ciceri, M. R., & Biassoni, F. (2015). How does a collision warning system shape driver's brake response time? the influence of expectancy and automation complacency on real-life emergency braking. *Accident Analysis & Prevention*, *77*, 72–81.

Ruskin, K. J., Corvin, C., Rice, S., Richards, G., Winter, S. R., & Ruskin, A. C. (2021). Alarms, alerts, and warnings in air traffic control: an analysis of reports from the aviation safety reporting system. *Transportation research interdisciplinary perspectives*, *12*, 100502.

Sarter, N. B., & Woods, D. D. (2017). Situation awareness: A critical but ill-defined phenomenon. *Situational awareness*, (pp. 445–458).

Saward, J. R., & Stanton, N. A. (2018). Individual latent error detection: Simply stop, look and listen. *Safety science*, *101*, 305–312.

Shams, M. Y., Elshewey, A. M., El-kenawy, E.-S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2023). Water quality prediction using machine learning models based on grid search method. *Multimedia Tools and Applications*, (pp. 1–28).

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *perception*, *28*, 1059–1074.

Simons, D. J., & Jensen, M. S. (2009). The effects of individual differences and task difficulty on inattentional blindness. *Psychonomic Bulletin & Review*, *16*, 398–403.

Thomas, L. C., & Wickens, C. D. (2001). Visual displays and cognitive tunneling: Frames of reference effects on spatial judgments and change detection. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 336–340). SAGE Publications Sage CA: Los Angeles, CA volume 45.

Varakin, D. A., Levin, D. T., & Fidler, R. (2004). Unseen and unaware: Implications of recent research on failures of visual awareness for human-computer interface design. *Human–Computer Interaction*, *19*, 389–422.

Wang, Y., Wu, Y., Chen, C., Wu, B., Ma, S., Wang, D., Li, H., & Yang, Z. (2022). Inattentional blindness in augmented reality head-up display-assisted driving. *International Journal of Human–Computer Interaction*, *38*, 837–850.

Werneke, J., & Vollrath, M. (2011). Signal evaluation environment: a new method for the design of peripheral in-vehicle warning signals. *Behavior research methods*, *43*, 537–547.

Wilkinson, R. T., & Seales, D. M. (1978). Eeg event-related potentials and signal detection. *Biological psychology*, *7*, 13–28.

Wu, X., & Li, Z. (2018). A review of alarm system design for advanced control rooms of nuclear power plants. *International Journal of Human–Computer Interaction*, *34*, 477–490.

Wu, Z., Zhao, L., Liu, G., Chai, J., Huang, J., & Ai, X. (2023). The effect of arhud takeover assistance types on driver situation awareness in highly automated driving: A 360-degree panorama experiment. *International Journal of Human–Computer Interaction*, (pp. 1–18).

Xu, J., Park, S. H., Zhang, X., & Hu, J. (2021). The improvement of road driving safety guided by visual inattentional blindness. *IEEE transactions on intelligent transportation systems*, *23*, 4972–4981.

Yang, Z., Shi, J., Zhang, Y., Wang, D., Li, H., Wu, C., Zhang, Y., & Wan, J. (2019). Head-up display graphic warning system facilitates simulated driving performance. *International Journal of Human–Computer Interaction*, *35*, 796–803.

Yu, X., Chen, C.-H., & Yang, H. (2023). Air traffic controllers' mental fatigue recognition: A multi-sensor information fusion-based deep learning approach. *Advanced Engineering Informatics*, *57*, 102123.

Zhou, F., Yang, X. J., & De Winter, J. C. (2021). Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE transactions on intelligent transportation systems*, *23*, 2284–2295.

About the Authors

Zhimin Li is currently a PhD candidate at the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. She received the MPhil degree in 2022 from the School of Management, Shenzhen University. Her research interest lies in ergonomics in air traffic control, and human-computer inter action.

Fan Li is an assistant professor of the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. She received the Ph.D. degree in School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. Her research interest lies in human-centered design, intelligent transportation systems, and sustainable human-computer interaction.

Mengtao Lyu is currently a PhD candidate at the Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University. His research interest lies in psychophysiological data-based engineering design, and human-system interactions. He got the MS.c degree in 2021 from the School of Mechanical and Aerospace Engineering, Nanyang Technological University.

Appendix

Table A1: The description and representation formats of the different warning types.

Warning types	Descriptions & Representation formats
Infringing radar separation minima	The planes infringe on radar separation minima if the vertical separation between planes is less than 1000 feet. For example: 1. Flight X and Flight Y are infringing radar separation minima. 2. Conflict: keep 1000 feet difference until both planes are on their localizer.
Infringing an area restriction	Certain regions within the airspace require a minimum flight altitude, and the violation of this altitude would trigger a warning. For example: Flight X is infringing an area restriction.
A plane entering	Upon entering the airspace, the plane triggers a warning to alert the ATCOs to its presence. For example: 1. Flight X heavy with you. 2. Flight X heavy, weather E. 3. Flight X heavy at FL70. 4. Flight X heavy is passing FL96 descending FL70. 5. Flight X heavy with information F.
Require further climb	Upon takeoff, the system alerts the ATCOs to instruct the plane to continue ascending, ensuring a successful departure from the airspace. For example: 1. Flight X requires further climb. 2. Flight X climbing FL60, BERGI departure. 3. Flight X passing 2100 climbing to FL60, BERGI departure. 4. Flight X passing 2100, BERGI departure.
Unable to land	A plane descending blindly to the glideslope from too steep an angle (over 3 degrees) or excessive altitude (over 4000 feet) will be warned that it is unable to land successfully. For example: 1. Flight X cannot capture the localizer at this heading. 2. Flight X is above the glideslope of runway 27. 3. Flight X is crossing the glideslope of runway 27. 4. Flight X is about to cross the localizer of runway 27.
Go around	If a plane exceeds 2000 feet in altitude or is within 4 miles of the preceding aircraft during approach, a go-around is required, as this concerns safe separation rather than the usual decision height of 1000 feet or below. For example: 1. Flight X is going around (too high). 2. Flight X is going around (too close behind a heavy). 3. Flight X is going around (runway not clear). 4. Flight X is going around (too fast). 5. Flight X is only 3.8 miles behind a heavy (need 4 miles). 6. Flight X is 0.2 miles too close to the preceding heavy aircraft.
Delays	A plane is delayed while it takes over 30 minutes to land. For example: Flight X is delayed.
About to leave the airspace	If the plane is overlooked and it is about to leave its landing airspace, the system will trigger a warning. For example: 1. Flight X is about to leave the airspace. 2. Flight X is almost leaving the airspace. 3. Flight X has just left the airspace.