SEEING THE UNSEEN THROUGH FIXATION DYNAMICS: LARGE LANGUAGE MODEL-AUGMENTED DETECTION FAILURE ANALYTICS IN AIR TRAFFIC CONTROL

LI ZHIMIN

PhD

The Hong Kong Polytechnic University

2026

The Hong Kong Polytechnic University

Department of Aeronautical and Aviation Engineering

Seeing the Unseen Through Fixation Dynamics: Large Language Model-Augmented Detection Failure Analytics in Air Traffic Control

LI Zhimin

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Aug. 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

__________________________ (Signed)

_________LI Zhimin____ (Name of student)

Abstract

Detection failure (DF) poses a critical challenge in air traffic control (ATC), where undetected alerts threaten decision-making and safety. Since DF is an unavoidable byproduct of human behavior in dynamic contexts, its accurate, proactive identification is vital, protecting thousands of lives under controllers' care. DF arises from inattentional blindness due to attentional constraints, yet encompasses broader perceptual and cognitive factors. Conventional methods rely on the presence or absence of fixations on the target to measure DF, but they overlook the cognitive continuum of visual processing, where fixation does not guarantee awareness, and human peripheral vision may detect target without direct fixations. This gap underscores the need for a structured DF classification that accounts for these divergent cognitive pathways and address complex "look-but-fail-to-see" and "seeing-without-looking" scenarios. Further complicating DF analysis is the spatiotemporal uncertainty of alerts in ATC, the dynamic nature of pre-alert attentional states, and class imbalance in DF datasets, which degrade the recall and F1-scores of current methods, necessitating a comprehensive analytical framework. This thesis proposes systematic DF analytics leveraging fixation dynamic-based feature representation and large language model (LLM)-augmented intelligent analysis. First, a gaze-behavioral DF classification grounded in Endsley's situational aware-

ness theory is developed, categorizing DFs into distinct types by awareness levels and identifying their key gaze indicators through a four-phase theoretical framework. Second, an ATC-specific hierarchical DF recognition framework is developed, integrating peripheral vision tracking, spatiotemporal uncertainty of alerts, and multitasking demands. Third, proactive DF prediction is achieved through a fixation-state transition feature set capturing pre-alert attention shifts, coupled with an LLM-driven probabilistic threshold tuning method to adaptively calibrate class-specific predictive probability boundaries.

Several ATC-specific case studies were conducted to validate the proposed methods. The results revealed significantly distinct gaze patterns across different DF types. The ATC-specific hierarchical DF recognition framework further identifies DF types through an Adaptive Symbolic Alert Detection method (95% precision in alert-region annotation) combined with LLM-based attention evaluation for reliable labeling, and a novel fixation coordinate-sensitive feature set that effectively captured spatiotemporal-frequency characteristics of diverse DF types (93% accuracy vs. 84% traditional features). Additionally, in proactive DF prediction, our proposed feature set outperforms existing features (78% vs. 71%–73% accuracy), with the LLM-enhanced prediction method further enhancing overall accuracy and significantly boosting DF-type-specific recall and F1-scores, showcasing superior robustness in class imbalance .

In conclusion, this research establishes a systematic paradigm for analyzing a broader and structured DF, introducing a streamlined feature set and a novel LLM-driven pathway for safety-critical systems. These developments lay the groundwork for designing intelligent, human-centered ATC support systems to improve alert detection reliability and aviation safety.

Publications arising from the thesis

Journal articles (in reverse chronological order)

[1] **Li, Z.**, Li, F., Xu, G., Li, D. "Beyond the Gaze: Peripheral Vision-Aware Visual Detection Failures Recognition Through LLM-Based Fixation Coordinate-Sensitive Analysis". In: *IEEE Transactions on Intelligent Transportation Systems* 99 (2025), pp. 1–20.

[2] **Li, Z.**, Li, F., Lyu, M. "Tracking the Unseen and Unaware: Deciphering Controllers' Detection Failures to Warnings Through Eye-Tracking Metrics". In: *International Journal of Human–Computer Interaction* (2025), pp. 1–20.

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

Manuscript under review

[1] **Li, Z.**, Lyu, M., Qiao, H., Yang, T., Li, F. "Predicting detection failures in air traffic control: leveraging fixation-state transitions and LLM-

driven probabilistic threshold tuning". In: *Advanced Engineering Informatics* (Aug. 2025).

Conference paper (in reverse chronological order)

[1] **Li, Z.**, Li, F. "Vigilant Air Traffic Control: Gaze-based Recognition of Detection Failures to Visual Warnings". In: *Advances in Human Factors of Transportation* 148.148 (2024).

[2] **Li, Z.**, Li, F., Lee, C.-h., Han, S. "Catching the Wanderer: Temporal and Visual Analysis of Mind Wandering in Digital Learning". In: *Engineering For Social Change*. IOS Press, 2024, pp. 729–738.

[3] **Li, Z.**, Li, Z., Li, F. "Visual Attention Analytics for Individual Perception Differences and Task Load-Induced Inattentional Blindness". In: *International Conference on Human-Computer Interaction*. Springer. 2023, pp. 71–83.

Acknowledgements

I would like to attribute the completion of this thesis to many people who have inspired me, encouraged me, and supported me.

First and foremost, I would like to express my deepest and most sincere gratitude to my supervisor, Dr. Li Fan, whose profound insights, patient guidance, and consistent encouragement have been instrumental in shaping not only my research capabilities but also the future trajectory of my life. She has inspired me to embrace new challenges with confidence and clarity. I am also deeply thankful to my co-supervisor, Dr. Xu Gangyan, for his invaluable advice and continuous encouragement throughout this journey. Their dedication has made this accomplishment possible.

My special appreciation goes to my master's supervisor, Prof. Yang Wen, whose enduring support and wisdom have served as a beacon of inspiration, guiding me to move forward with calm and courage. I am also grateful to my advisors—Prof. Chen Xingyu, Dr. Han Su, Dr. Lyu Mengtao, Dr. Li Ruilin, and Dr. Cui Jian—for their generous guidance and expertise, which greatly enriched my research experience.

I also wish to express my sincere gratitude to my teammate-siblings in the "SHINE group" and my friends who have accompanied me throughout my PhD studies. Our sincere friendship, thoughtful discussions, and joyful companionship have made my research experience truly memorable.

Lastly, I extend my utmost appreciation to my beloved parents and boyfriend, whose unconditional love, endless understanding, and consistent support have been my foundation and source of strength throughout this meaningful endeavor.

Contents

List of Figures

List of Tables

Nomenclature

ADASYN	Adaptive Synthetic Sampling
AI	Artificial Intelligence
AOI	Area of Interest
ATCO	Air Traffic Control Officer
BF	Blink Frequency
BOHB	Bayesian Optimization Hyperband
BO-TPE	Bayesian Optimization with Tree-structured Parzen Estimator
CD	Correct Detection
DF	Detection Failure
ECG	Electrocardiogram
ED	Effective Detection
EEG	Electroencephalogram
FC	Fixation Count
FD	Fixation Duration
FVT	First View Time
GBDT	Gradient Boosting Decision Tree
GNN	Graph Neural Network
HCI	Human-Computer Interaction

IB	Inattentional Blindness
kNN	k-Nearest Neighbors
LBFTS	Look-but-fail-to-see Errors
LightGBM	Light Gradient Boosting Machine
LLM	Large Language Model
L-PTT	LLM-driven Probabilistic Threshold Tuning
MI	Misinterpretation
MLP	Multilayer Perceptron
MPD	Mean Pupil Diameter
MLPD	Mean Left Pupil Diameter
MRPD	Mean Right Pupil Diameter
MSA	Mean Saccade Amplitude
NDF	Non-Detection Failures
OB	Ordinary Blindness
PCA	Principal Component Analysis
PD	Pre-attentive Detection
PR	Precision-Recall
PVN	Peripheral Vision Neglect
RF	Random Forest
ROC	Receiver Operating Characteristic
SAM	Segment Anything Model
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine

XGBoost	eXtreme Gradient Boosting

Chapter 1

Introduction

This introductory chapter is organized into five sections. Section 1.1 begins by presenting the research background on human detection failures in aviation, highlighting the value of visual attention analysis in this context and motivating the study. Section 1.2 states the research objectives formulated from this background. The scope of the research is defined in Section 1.3, while Section 1.4 elaborates on its significance. Finally, Section 1.5 provides an overview of the thesis structure.

1.1 Background

Conflict detection and resolution are critical tasks in air traffic control (ATC), requiring air traffic controllers (ATCOs) to maintain a high level of vigilance and make precise decisions in complex and dynamic environments [1]. As global air traffic density increases and automation systems become more prevalent, ATCOs face growing challenges in monitoring digital interfaces, significantly elevating risks of *detection failure* (DF) – a type of human error where operators fail to notice or respond to unexpected and salient stimuli or signals [2]. Within high-stakes ATC operations, DF constitutes a primary contributor to operational incidents. Empirical evidence demonstrates their significant impact: research reported visual-perceptual failures as contributing to nearly 70% of all perceptual errors in 48 instances of loss of separation incidents, with missed or delayed detection of critical warnings being particularly prevalent [3]. Additionally, U.S. Federal Aviation Administration (FAA) 2022 data documented 1,732 runway incursions, many attributable to undetected visual cues [4, 5]. A striking example includes the 2024 Haneda Airport collision in Japan, where ATCOs failed to detect a flashing warning signal, resulting in a catastrophic runway collision [6]. These incidents collectively underscore the urgent need for systematic DF analytics to mitigate operational risks in ATC environments.

DF, as a category of human error rooted in cognitive psychology, is closely related to inattentional blindness (IB), where attentional limitations prevent individuals from noticing salient visual stimuli [7, 8]. Foundational studies by Simons and Chabris [9, 10] and Mack and Rock [11, 12] provided critical insights into the mechanisms of IB, establishing a theoretical basis for DF research. However, while IB primarily stems from attentional allocation constraints, DF encompasses broader cognitive and perceptual limitations, such as perceptual constraints, cognitive biases, and expertise-driven limitations, with IB representing only one manifestation [13–15]. The absence of a comprehensive theoretical framework to systematically classify these diverse DF mechanisms leaves its underlying pathways poorly understood, particularly in complex settings like ATC . On the other hand, traditional methods for identifying DF—such as self-reporting and retrospective analysis—are limited by their subjective nature and inherent inability to reliably capture failures in dynamic operational settings [5, 11]. In recent years, eye-tracking technology has emerged as an objective tool for DF investigation, enabling the analysis of physiological indicators such as fixation patterns and scanpaths to better understand monitoring behaviors and attention lapses [16–18].

Therefore, this study aims to advance the analysis of DF in ATC by developing a comprehensive framework to address critical shortcomings identified in the process. Through this analysis, three pivotal gaps have emerged that need to be resolved to enhance DF understanding and analytics within the ATC context, as outlined below:

- **The absence of a systematic and theoretical framework for DF**: Few systematic theoretical models have been developed to differentiate the varied cognitive and perceptual mechanisms underlying DF in ATC, hindering precise identification and intervention [3]. For instance, "look-but-fail-to-see" (LBFTS) errors, where ATCOs fixate on a warning yet fail to process it cognitively, remain indistinguishable from complete misses using conventional eye-tracking metrics [12, 15]. This gap reflects an overreliance on gaze-based detection without integrating broader situational awareness theories, limiting the ability to identify a broader range of DF types, such as those driven by perceptual overload or expertise-driven biases.
- **The lack of a DF recognition framework adapted to the ATC domain:** ATCOs operate in visually complex settings where warnings appear unpredictably across multiple displays (e.g., radar, auxiliary panels) [3, 19]. Traditional eye-tracking analyses rely on static Areas of Interest (AOIs)—predefined zones on a screen assumed to contain critical information [20,

21]. These fixed boundaries cannot dynamically adjust to the real-time movement of alerts across displays or account for ATCO's shifting fixations, making it struggle to adapt to this spatiotemporal variability of ATC-specific alerts [22]. Moreover, ATCOs rely on peripheral vision during multitasking, yet most studies focus on central eye movements, ignoring subconscious tracking and scanning behaviors [23, 24]. This gap limits the accuracy of DF recognition in ATC context, necessitating new approaches to effectively capture peripheral visual dynamics.

- **Unaddressed proactive DF prediction in ATC:** In the field of aviation safety, the transition from passive identification to proactive prediction of DF is of critical importance [25–27]. This shift not only pertains to the timeliness of error analysis but also directly determines the effective window for implementing preventive measures [28]. In the development of next-generation intelligent control systems, there is a pressing need for systems to possess the capability to understand controllers' real-time attention states, rather than merely responding to operational commands [29]. However, DF prediction remains unexplored mainly because DF events are rare compared to normal operations - creating a fundamental class imbalance challenge – and traditional eye-tracking metrics struggle to reliably detect the subtle changes in attention states before DF occurs [25]. This absence of DF prediction limits the potential for elevating ATC safety standards.

This study introduces a rigorous analysis of DF through a systematic theoretical framework, categorizing DF types, examining their root causes and eye-tracking indicators in ATC environments, and developing methods for DF recognition and prediction.

1.2 Objectives

The objectives of this research are aligned with addressing the research gaps associated with DF analytics in ATC, focusing on understanding, recognizing, and predicting DFs within the context of ATCOs' operational environments, as shown in Figure 1.1. This goal is separated into three sub-objectives:

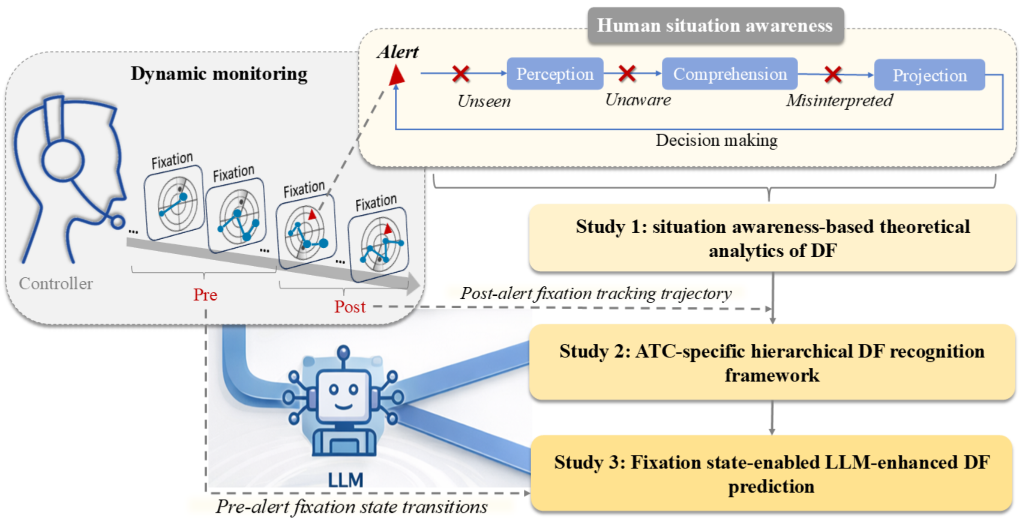


Figure 1.1: Organization of the thesis

- *Conceptualize and categorize DF*

  Theoretically classifying DF based on Endsley's situation awareness theory is the first step in DF analytics [30]. According to Endsley's theory, warning detection involves three stages: perception (detecting a warning), comprehension (understanding the warning), and projection (anticipating potential risks) [13, 31]. Disruptions at each stage may lead to different types of DF.

This objective aims to categorize DF based on varying situation awareness levels, exploring their underlying mechanisms, specific causes, and key visual features to deepen understanding of how visual and cognitive factors jointly affect alert detection performance, while evaluating how different warning frequencies influence these DF patterns.

- *Recognize DF in dynamic ATC contexts*

  Based on the DF categorization, this objective aims to design a DF recognition framework specific to ATC environments that accounts for the spatiotemporal uncertainty of alerts and peripheral vision tracking in multitasking [3]. This framework aims to effectively identify alert regions appearing at uncertain times and locations, intelligently assess controllers' visual engagement with these alerts via large language model (LLM), and recognizing DF types by accurately characterizing their dynamic fixation trajectories, thereby better capturing their scope and patterns of visual attention allocation.

- *Predict DF through fixation state-enabled LLM analysis*

  This objective aims to predict DF through streamlined feature extraction and adaptive modeling techniques. It intends to encode temporal attentional switching patterns using fixation-state transitions [32, 33]. Additionally, it aims to optimize probabilistic thresholds for diverse DF types to address the class imbalance issues [34]. In contrast to static threshold approaches, the LLM-driven threshold tuning method seeks to derive adaptive, class-specific probability thresholds to improve prediction performance.

1.3 Scope

Concerning the research objectives, this work conducts DF analytics from three aspects, including theoretical understanding, operational recognition, and predictive modeling of DFs in ATC systems. The scope of the research is as follows:

- *A Theoretical framework to conceptualize DF*

  This study focuses on developing a theoretical framework to conceptualize DF, specifically targeting visual DF related to ATC operational warnings, such as separation breaches, airspace violations, and procedural directives. The scope encompasses the application of situation awareness theory to systematically analyze the formation pathways of DF, proposing a gaze-behavioral DF classification framework to categorize these errors based on visual-behavioral patterns [30]. The DF classification is validated by a controlled DF induction experiment involving a basic ATC monitoring task designed to induce diverse DFs and collect eye movement data. The experiment examines gaze dynamics across continuous and intermittent warning frequencies to understand attentional patterns. The scope further extends to DF-type recognition and analytics, employing Shapley Additive Explanations to interpret model outcomes and assess feature importance [25]. Excluded from this study are non-visual DFs and warnings not related to basic ATC operational tasks, ensuring a focused investigation into visual-based DF mechanisms in ATC environments.

- *An ATC-specific hierarchical DF recognition framework*

  This study focuses on developing a hierarchical framework for recognizing DF specific to the ATC domain, emphasizing visual attention dynamics through eye-movement analysis. The scope encompasses the integration of peripheral vision tracking during multitasking, assessing central vision engagement by evaluating fixation trajectory overlaps with dynamic alert regions, rather than relying on a fixed 2° central vision range, to account for the variable size and position of ATC alerts [24]. The framework is validated through ATC simulation experiments featuring spatiotemporally uncertain radar-based warnings, excluding fixed alert regions. An Adaptive Symbolic Alert Detection method is included to dynamically identify and annotate radar alerts, supplemented using GPT-4V LLMs to automate the evaluation of controllers' visual attention, addressing the challenges of manual labeling. The study further develops a fixation coordinate-sensitive feature set, analyzing fixation trajectories across spatial, frequency, and time domains. The scope is limited to eye-movement-based DF recognition, explicitly excluding other physiological signals (e.g., heart rate or EEG) and non-ATC contexts, ensuring a targeted investigation into visual attention dynamics in ATC environments.

- *A Fixation State-enabled LLM-Enhanced DF Predictor*

  This study focuses on developing a fixation state-enabled, LLM-enhanced predictor for DF in ATC, targeting the proactive prediction of diverse DF types. The scope includes designing a fixation-state transition feature set to capture pre-alert attentional switching patterns, implementing a novel LLM-driven Probabilistic Threshold Tuning (L-PTT) method to dynamically adjust class-specific thresholds for improved sensitivity to critical DF events, and conducting experiments to evaluate input feature formats (2D heatmaps versus numerical tables) and tuning strategies (exhaustive versus priority-based) to optimize LLM integration. This targeted scope supports the creation of intelligent, human-centered ATC systems by leveraging fixation data and LLM for adaptive DF prediction.

1.4 Significance

By achieving these goals, this study will advance the following research domains:

- *Exploratory research on DF analytics in ATC operations*

  While practitioners have paid increasing attention to the criticality of DF in ATC, the exploration of systematic approaches to DF analytics remains underdeveloped. This research positions itself as a pioneering effort to investigate the theoretical underpinnings of DF mechanisms, particularly through eye-tracking measurements in ATC environments. By developing streamlined fixation-specific metrics—including fixation-coordinate dynamics during alert periods and fixation-state transition features in pre-alert phases—this study lays a groundwork for DF analytics in ATC operations, opening a new paradigm for optimizing ATC alert detection reliability and enhancing aviation safety.

- *Development of LLM-augmented analytics for adaptive DF management in ATC*

  This research expands the application scope of LLMs by integrating LLM-driven analytics to provide adaptive insights into DF management in ATC.

Key advancements include intelligent processing of visual engagement in alert regions, dynamic learning of relationships between eye movement features and predictive probabilities, and adaptive probabilistic threshold tuning for DF prediction. This work establishes a groundbreaking LLM-enhanced research paradigm that bridges human attentional dynamics with AI-driven decision-making. By demonstrating LLMs' potential in safety-critical ATC operations, the study creates a transformative framework extendable to other high-stakes domains like autonomous systems and medical diagnostics.

1.5 Structure of the thesis

This thesis comprises six chapters, as illustrated in Figure 1.2.

Chapter 1 introduces the research background, objectives, scope, significance, and overall structure.

Chapter 2 reviews DF and eye-tracking studies in aviation, discusses class-imbalance solutions, and identifies research challenges in developing intelligent ATC support systems. It also examines potential applications of LLMs for DF analysis.

Chapter 3 establishes a four-phase DF classification framework and identifies gaze-based indicators. An experimental study validates the proposed model and examines the effect of warning frequency on DFs, revealing distinct gaze patterns and achieving robust four-category recognition.

Chapter 4 presents a hierarchical DF recognition framework integrating peripheral vision and attentional states. It introduces an ATC-specific alert detection method and an LLM-based visual attention evaluation, achieving high accuracy through a novel fixation-aware feature set. A case study validates the framework.

Chapter 5 proposes a predictive approach using fixation-state transition features and LLM-based L-PTT method for improved DF prediction. It demonstrates the advantage of 2D heatmaps over numerical data for feature interpretation and evaluates threshold strategies through a case study with expert validation.

Chapter 6 summarizes the research contributions, discusses limitations, and suggests future research directions.

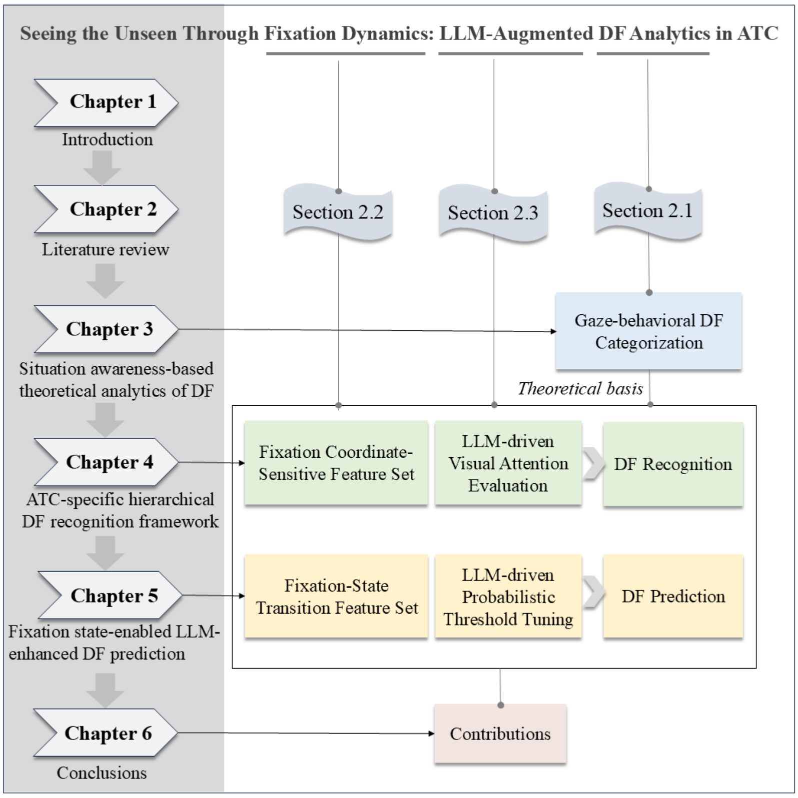


Figure 1.2: Organization of the thesis.

Chapter 2

Literature Review

In this chapter, a comprehensive review of the literature on DF in ATC is conducted, with a focus on cognitive foundations, eye-tracking analytics, and computational solutions. The chapter is structured as follows: Section 2.1 examines the cognitive and behavioral underpinnings of DF, including conceptual analyses, empirical approaches, alert detection studies, and identifies prevailing research gaps. Section 2.2 explores eye movement measurements and their application in DF analytics, highlighting relevant metrics and persistent challenges. Section 2.3 discusses machine learning solutions for class-imbalanced DF analysis, covering recognition models, imbalance mitigation methods, potential LLM applications, and remaining challenges in LLM-augmented DF analysis.

2.1 Cognitive and behavioral foundations of DF

This section explores the cognitive and behavioral underpinnings of DF in ATC, focusing on related concepts such as IB, change blindness, and mind wandering. It clarifies their definitions, relationships to DF, and their implications for attentional lapses in high-stakes environments.

2.1.1 Conceptual analysis of DF

DF in ATC refers to the failure to detect or respond to critical cues, such as radar alerts, compromising safety. Although DF has not been extensively studied as a standalone phenomenon in ATC, related cognitive and perceptual failures have been extensively explored in cognitive psychology and human factors research. As a human error rooted in cognitive psychology, DF is closely related to IB, where individuals miss salient, unexpected stimuli due to attentional limitations [9, 11]. However, DF encompasses broader cognitive and perceptual limitations, and its mechanisms remain underexplored [35, 36]. This section traces the historical development of the related phenomena to clarify their interconnections and contributions to understanding DF. It then evaluates theoretical frameworks to assess their applicability to DF in ATC's dynamic context, highlighting the comprehensive fit of Endsley's situation awareness model.

The understanding of DF builds on a historical progression of research into attentional and perceptual failures, which collectively informs its analysis in ATC. Early studies on selective attention showed that cognitive resources prioritize task-relevant cues, often at the expense of peripheral information [37]. This laid the groundwork for IB, where focused attention causes individuals to miss salient, unexpected stimuli, such as ATC alerts during aircraft monitoring [9, 11]. IB's close alignment with DF, due to the unexpected nature of alerts, evolved into research on change blindness, where attentional disruptions prevent detection of visual scene changes, like dynamic radar updates [38]. This progressed to looked-but-failed-to-see errors, where controllers scan but fail to register critical cues due to cognitive overload or misprioritization, directly contributing to DF [3, 15]. Subsequent research identified mind wandering, where attention drifts to internal thoughts during prolonged shifts, and attentional tunneling, where excessive focus on one information source neglects others, both exacerbating DF in ATC's high-pressure environment [39, 40]. Complementing these, pre-attentive detection via peripheral vision describes unconscious processing of stimuli without direct fixation, driven by covert attention [23]. This automatic detection can mitigate DF by facilitating reflexive responses but may fail under cognitive overload, contributing to DF's complexity. Collectively, these phenomena—termed degraded states of engagement by [39]—highlight a progression from general attentional filtering to specific failures in dynamic, safety-critical environments, providing a conceptual foundation for analyzing DF in ATC.

Theoretical frameworks provide structured approaches for linking these phenomena to model DF's cognitive underpinnings in ATC. Signal detection theory, developed by [41], models detection tasks as decisions under uncertainty, where DF results from high detection thresholds or low sensitivity to signals amidst noise. In ATC, SDT explains how controllers may fail to classify radar signals as critical due to workload or ambiguous cues [35]. Visual information processing models, which describe how sensory input is filtered and prioritized, highlight the role of pre-attentive visual selection in DF, where irrelevant stimuli may overshadow critical ones [37]. Similarly, the theory of early visual processing [42] posits that early visual mechanisms optimize information encoding, but failures in this process, such as inefficient feature extraction, can contribute to DF in complex visual fields like ATC displays. While these models offer valuable insights, Endsley's situation awareness model [43] is particularly suited for ATC due to its focus on dynamic, multi-level cognitive processing. The SA model delineates three levels—perception (Level 1), comprehension (Level 2), and projection (Level 3)—and frames DF as a breakdown in perceiving cues (e.g., due to IB or failed pre-attentive detection), comprehending their significance (e.g., looked-but-failed-to-see errors), or anticipating future states (e.g., attentional tunneling or misdetection) [14, 44]. Unlike SDT, which focuses narrowly on detection thresholds, or visual processing models, which emphasize low-level perceptual mechanisms, the SA model captures the holistic cognitive demands of ATC, where controllers must integrate multiple information sources under time pressure [13]. This makes it uniquely effective for analyzing DF's cognitive roots and guiding interventions in safety-critical contexts.

By tracing the historical development of attentional and perceptual failures and grounding them in theoretical models, this section establishes a robust framework for understanding DF in ATC. IB, pre-attentive detection, and related phenomena provide multifaceted mechanisms for analyzing DF's cognitive roots, while Endsley's SA model stands out for its ability to integrate these phenomena into a cohesive model of cognitive performance in dynamic environments, offering a foundation for subsequent analyses of eye movement metrics and predictive modeling in ATC.

2.1.2 Empirical approaches to analyze DF

Early studies on DF in ATC relied on self-reporting and retrospective analyses to identify perceptual and cognitive failures, including IB and other detection errors [11]. For example, Shorrock interviewed 28 ATC controllers and analyzed 48 aircraft proximity incidents, classifying errors like late detection and misdetection through subjective accounts, providing a foundational taxonomy for DF [3]. Additionally, post-experiment questionnaires were used in a simulated flight task to identify failures to notice objects during distracting tasks [5], highlighting the role of divided attention in DF. Similarly, targeted questionnaires were employed in driving scenarios to assess failures to detect incongruent objects, linking task relevance to detection errors [45]. Subjective reports with behavioral metrics were supplemented, observing failures to perform operations in ATC notifications as indicators of DF [35]. Load theory [46] was tested in a subitizing task, using performance metrics to explore how excessive perceptual load leads to DFs, relevant to ATC's high-cognitive-demand environment [47]. Moreover, the research team of Anderson investigated habituation to security warnings, using questionnaires and behavioral responses to show how repeated exposure reduced detection rates [48, 49]. These methods, while critical for establishing DF's cognitive underpinnings, were limited by recall biases and the inability to capture dynamic perceptual processes, necessitating more objective approaches.

The limitations of subjective methods prompted a shift to objective physiological measurements, with eye-tracking emerging as a practical, non-invasive tool widely adopted for studying attentional processes in DF. Eye-tracking was applied in simulated aviation tasks to confirm DFs, measuring fixation patterns to complement questionnaire-based findings [50]. For instance, longer fixation duration were demonstrated to be associated with stimuli detection, highlighting eye-tracking's ability to quantify attention allocation [51]. In sustained detection tasks, research revealed dissociations between looking and perceiving, critical for understanding DF in ATC [52]. Blink rate patterns was linked to engagement levels, offering a measurable indicator of attentional lapses relevant to DF [53]. Other physiological methods, such as EEG, have also been explored in contexts like cybersecurity [49], ATC stress [54, 55], and mind-wandering [56], but eye-tracking's non-intrusive nature and ease of application in ATC settings have made it a preferred method [20]. Given its practical advantages and widespread use, this study focuses on eye-tracking to analyze DF's attentional mechanisms, as detailed in subsequent sections.

2.1.3 Studies on alert detection in ATC

Research on alert detection in ATC focuses on optimizing visual warning designs, addressing auditory warning annoyance, and evaluating warning frequency to mitigate DF and enhance situational awareness. Existing research emphasized that visual alerts, such as radar display signals, are critical for conveying hazard-specific information in ATC's dynamic environment, requiring clear and consistent designs to ensure effective attention allocation [36]. However, frequent visual signals like Conflict Alerts (CA) and Minimum Safe Altitude Warnings (MSAWs)are often missed during high workload [19], contributing to DF when excessive alerts erode trust, as seen in the "cry wolf" effect [57]. Auditory alerts exacerbate this issue, with Friedman-Berg et al. (2008) reporting that 62% of CAs and 91% of MSAWs in en route settings are perceived as unnecessary, causing annoyance and desensitization that impair responsiveness [22]. It is demonstrated that semantic visual alerts, providing context-specific information, significantly reduce response times and enhance situational awareness for critical events like Short Term Conflict Alerts, outperforming annoying acoustic alerts [58, 59]. Proactive visual warning systems in virtual environments improve hazard detection without disrupting task performance, suggesting potential for real-time ATC applications [60]. Related studies on warning frequency highlight its impact, with frequent warnings linked to information overload and reduced detection rates [61], and incomplete warning information impairing decision-making in control settings [62]. Excessive warnings can also erode trust, leading operators to deprioritize alerts [63]. These findings support the need for further research on the impact of continuous versus selective warning strategies on DF and situational awareness, underscoring warning frequency as a critical direction for designing human-centered visual alerts to optimize attention and minimize auditory-induced annoyance in ATC's safety-critical context.

2.1.4 Remaining gaps in DF research

A comprehensive theoretical framework of DF remains understudied, lacking systematically classify its diverse error mechanisms. Objective analysis methods, like eye-tracking, have shown promise for detecting attentional lapses but are rarely applied to study DF specifically in ATC's high-stakes environment. Furthermore, excessive warnings also erode operator trust, exacerbating DF risks, yet effects of dynamic alert designs adjusting frequency on DF occurrences are underdeveloped. Integrating eye-tracking with warning frequency analysis to examine DF in the context of ATC alerts is a critical gap, offering a vital direction for developing human-centered, objective approaches to DF analytics.

2.2 Eye movement measurements and insights into DF analytics

Eye-tracking is a critical tool for studying attentional lapses relevant to DF in ATC, where controllers may miss vital cues like radar alerts, risking safety. Although direct eye-tracking studies on DF in ATC are scarce, research on related attentional processes—such as IB, perceptual overload, fatigue, and distraction—offers valuable insights [64]. This section synthesizes how eye-tracking metrics, particularly fixation duration, frequency, and scanpath trajectories, have been applied to these contexts, and discusses challenges in adapting these approaches to ATC's dynamic environment.

2.2.1 Eye-tracking analyses relevant to DF

Eye-tracking research, while limited in directly addressing DF in ATC, provides critical insights into attentional lapses such as IB, high workload, and fatigue, which are relevant to missed detections in safety-critical settings [28]. Early studies on IB demonstrated that fixation patterns are key indicators of DFs, as individuals may look at stimuli without perceiving them [65]. It is found that both noticers and non-noticers fixated on unexpected objects in dynamic tasks, but only those with sufficient cognitive processing registered them, suggesting that fixation alone does not guarantee detection, a critical concern for ATC controllers scanning radar displays [52]. Similarly, a study showed that stimuli crossing the fovea were missed without adequate fixation duration, emphasizing the need for sustained attention to critical alert zones in ATC [51]. Building on this, Krueger et al. revealed that microsaccade rates decrease under high visual attention demands, distinguishing “looking” from “seeing,” which could help identify DF events where controllers fail to process alerts despite gazing at them [66].

These findings extend to workload and fatigue, which further exacerbate attentional lapses in ATC’s demanding environment. Ranti et al. linked reduced blink rates to higher engagement with visual content, while increased blink duration signaled disengagement in prolonged tasks [53]. High workload compounds these issues, as shown by [67], who used spatial density and fixation duration to demonstrate that expert users focus more efficiently on critical areas in electronic medical record tasks, a principle applicable to optimizing ATC display monitoring. Eye-tracking metrics like blink behavior and head movements outperformed vehicle-based features in detecting driver drowsiness, suggesting that direct monitoring of fixation patterns could enhance DF detection [68]. In complex visual tasks, Krejtz et al. used scanpath aggregation and transition matrix entropy to reveal focal attention in experts, contrasting with ambient scanpaths in non-experts, highlighting the need for consistent scanpaths to maintain situation awareness [69]. Similarly, it is showed that fixation patterns predict lane-change intentions in driving, indicating that guided saccades could improve attention to critical cues in multi-screen interfaces [70]. These studies collectively underscore fixation duration, frequency, saccades, scanpaths, and blinks, often analyzed via Areas of Interest (AOI), as pivotal for detecting attentional lapses. By modeling these metrics, eye-tracking can inform strategies to mitigate DF in ATC, supporting real-time monitoring and adaptive interventions for safety-critical tasks.

To anticipate DF in ATC, eye-tracking studies highlight fixation duration as the primary metric for assessing attention states, with saccades and pupil diameter providing complementary insights into cognitive readiness for detecting critical signals [71]. Specifically, Holmqvist et al. emphasized that fixation durations, typically ranging from 200–500 ms in tasks like reading and scene perception, reflect cognitive processing depth, with shorter durations (100–200 ms) indicating rapid scanning of familiar stimuli, and longer ones (over 500 ms) signaling complex processing [72]. In addition, research found that fixation durations in reading (200–400 ms) increase to over 500 ms for complex texts, suggesting that prolonged fixations before an alert may indicate cognitive overload [73]. In scene perception, Pannasch et al. [74] noted that shorter fixations (100–200 ms) correlate with exploratory scanning, while longer fixations (300–500 ms) reflect focal attention, implying that pre-alert fixation patterns could predict whether controllers are primed to detect alerts. Negi et al. [32] further showed that in subtitle learning, fixation durations shift from 100–200 ms for quick scans to over 500 ms during confusion, with saccade amplitude and pupil dilation indicating shifts in attention intensity. These metrics, analyzed via AOI, suggest that shorter fixations and reduced pupil dilation before alerts may indicate low engagement, while prolonged fixations and larger saccades reflect heightened attention, offering a pathway to anticipate DF in ATC by monitoring pre-alert attention states.

2.2.2 Remaining challenges in eye-tracking for DF analysis

Current eye-tracking research often focuses on specific attentional lapses, such as those due to IB, workload or fatigue, using metrics like fixation duration or blink rate, but distinguishing multiple attentional lapses underlying DF in a unified classification system remains underexplored. Developing a streamlined feature set of eye movement metrics to differentiate diverse DF causes is challenging yet critical for practical ATC applications [75]. Additionally, traditional AOI-based analyses, as used in [20, 52, 76], rely on fixed regions like alert zones, which are less effective in ATC's dynamic, spatiotemporally uncertain alert patterns [3]. Static AOI methods struggle to capture attention in unpredictable interfaces, necessitating more effective metrics focusing on visual tracking and attention scope to enhance DF analysis in ATC's complex environment.

2.3 Solutions for class-imbalanced DF analysis using LLMs

This section reviews models for DF recognition and prediction, methods to address class imbalance in DF datasets, potential applications of LLMs for DF analysis and imbalance mitigation inspired by recent literature, and remaining challenges, particularly in leveraging LLMs for adaptive DF analysis in ATC's dynamic environment.

2.3.1 Models for DF recognition and prediction

Recent studies have employed a range of machine learning and deep learning models to analyze eye-tracking data, focusing on metrics like fixation patterns, blink behavior, and gaze dynamics to identify these lapses. For instance, convolutional neural network-based approaches have been used to predict eye fixation locations in dynamic environments like driving, which is relevant to modeling scanpath trajectories on ATC radar displays [27, 65]. However, such models primarily predict eye fixations and are less suited for directly identifying DF types, as they focus on intermediate eye movements rather than attentional outcomes.

In contrast, models designed for the classification of attentional states or task performance have shown robust applicability for recognition and prediction. These models leverage eye-tracking features to detect cognitive states, such as fatigue or loss of situation awareness [26, 76, 77]. For example, support vector machines have been applied to identify fatigue-related lapses by analyzing blink behavior and head movements in driving contexts [68]. Similarly, ensemble methods like random forest have been used to predict stress-related performance lapses in industrial tasks, integrating eye-tracking with neurophysiological data [78, 79]. Advanced architectures, such as Transformer-based models combined with multilayer perceptrons, have also emerged for classifying cognitive states like reading comprehension, offering potential for DF analysis by modeling multimodal eye-tracking and EEG data [80–82]. Additionally, specialized models like LSTM with attention mechanisms and convolutional networks have been developed to capture structured gaze dependencies and local dynamics, enhancing eye movement recognition relevant to DF [83, 84]. While a variety of models, including support vector machines, gradient boosting methods, neural networks, and Transformer-based architectures, have been identified as effective, their selection depends on the specific DF mechanism and task context, underscoring the need for tailored approaches in ATC's dynamic environment.

2.3.2 Methods for addressing class imbalance

Class imbalance, where DF events are rarer than routine successful detections, poses a significant challenge for developing robust DF analysis models. Current approaches to mitigate this issue can be categorized into three main areas: data-level methods, algorithmic methods, and post-processing techniques, each addressing the skewed distribution of DF events in distinct ways. Data-level methods which modify the dataset to balance class representation, are widely used as a basic pre-processing step, with oversampling techniques being particularly common. Oversampling methods like Synthetic Minority Over-sampling Technique (SMOTE) [85] generate synthetic minority class samples to balance datasets, while Adaptive Synthetic Sampling (ADASYN) [86] focuses on creating samples near challenging boundary instances, improving sensitivity to rare DF events. Brandt and Lanzén [87] compared SMOTE and ADASYN across imbalanced datasets, finding that both improve classifier performance, though SMOTE often outperforms ADASYN with support vector machines as class imbalance increases, while both enhance random forest performance similarly. Undersampling, which reduces majority class samples, and ensemble-based sampling methods like EasyEnsemble [88], which use balanced bootstrap sampling to create multiple balanced subsets for training, are also employed, though less frequently than oversampling due to potential information loss. Algorithmic methods adjust model training to prioritize minority classes. For instance, class-weighted objectives [89] modify loss functions to assign higher penalties to misclassifications of rare DF events, enhancing detection accuracy. Additionally, Bayesian optimization techniques are introduced [90], such as Bayesian Optimization with HyperBand (BOHB) [91] and Bayesian Optimization with Tree-structured Parzen Estimator (BO-TPE) [92], which combine Bayesian modeling with early-stopping and tree-structured density estimation to optimize hyperparameters for imbalanced datasets, improving model performance on rare events.

Post-processing techniques play a critical role in optimizing model outputs for imbalanced DF datasets, particularly through threshold adjustment methods, which serve as a key baseline for detecting rare events like missed alerts in ATC. Probability calibration [93] adjusts predicted probabilities to align with true class distributions, improving the reliability of DF detection. Threshold adjustment using Receiver Operating Characteristic (ROC) / Precision-Recall (PR) analysis [94] or grid search [34] optimizes metrics like F1-score or Youden's index, enabling precise identification of DF events despite their rarity. These static methods are essential for establishing robust performance benchmarks to enhance DF analysis in ATC's dynamic environment.

These approaches can enhance the ability to model rare DF events, but their effectiveness depends on dataset characteristics and task context. The choice of methods requires careful consideration of trade-offs, such as information loss in undersampling, with post-processing techniques providing a critical foundation for optimizing detection performance.

2.3.3 Potential applications of LLM for DF analysis and class imbalance

Although LLMs have not been applied to DF analysis or class imbalance issues, recent studies in related domains suggest their potential for analyzing complex eye-tracking patterns and providing decision-making support [29]. For instance, LLMs can process images and leveraging their contextual reasoning to assign labels, as shown in chest X-ray analysis where LLMs combined with gaze heatmaps improved diagnostic accuracy [95, 96]. This ability extends to integrating multimodal data, such as eye-tracking and EEG, to classify cognitive states, with studies achieving over 60% accuracy in word-level neural state classification during reading comprehension [80, 81]. Similarly, LLMs have been used to synthesize conditioned traffic data for autonomous driving simulations [97] and accident investigation [98]. In task-oriented contexts, LLMs have simulated human-like coordination, integrating gaze and situational data to predict behavioral outcomes, adaptable to ATC's dynamic environment [99]. Furthermore, LLMs support decision-making by generating actionable recommendations, such as dynamic word cards for communication based on gaze patterns [100] or context-aware pilot monitoring to prevent out-of-the-loop states [20]. These capabilities are complemented by LLMs' ability to provide fine-grained reasoning, as seen in [80, 101] and medical diagnosis with knowledge-based prompts [102]. For instance, LLM-powered robots have shown promise in enhancing human-robot interaction by adapting to gaze-driven cognitive states [103]. By leveraging these strengths, LLMs could enable adaptive threshold adjustment for class imbalance through modeling relationships between sparse data points and probabilistic outcomes, though practical

ATC applications remain unexplored, highlighting a critical research gap.

2.3.4 Remaining challenges in class-imbalanced DF analysis

Despite advances, significant challenges persist in applying models and LLMs to DF analysis in ATC. The absence of LLM-based approaches for DF analysis and class imbalance is a critical gap, as their contextual modeling capabilities could enable adaptive threshold adjustment for rare DF events, yet few studies have explored this in ATC. Dynamic threshold adaptation remains challenging, as static methods like ROC analysis [94] struggle with relationship learning between eye movement patterns and predictive probability, requiring adaptive probability adjustments. Developing an intelligent predictor leveraging LLMs for adaptive eye movement data analysis, remains a critical area for in-depth investigation in ATC safety.

Chapter 3

Situation awareness-based theoretical analytics of DF

In this chapter, a situation awareness-based theoretical framework is proposed to analyze DF in ATC. The study establishes a gaze-behavioral DF classification and examines how warning frequency influences operators' visual patterns and error types. The structure of this chapter is organized as follows: Section 3.1 presents the research motivation and identifies key challenges in DF analytics. Section 3.2 introduces the methodology, including the proposed DF classification, experimental design for DF induction, gaze data collection procedures, and analytical approaches for recognizing DF types. Section 3.3 presents the results, highlighting gaze dynamics across DF types and warning frequencies, recognition performance, and feature importance analysis using SHAP. Finally, Section 3.4 discusses the implications of these findings.

3.1 Background

3.1.1 Motivation

Numerous ATC incidents arise from overlooking potential safety risks due to operators' focused awareness being limited to only a small subset of the visual environment [2]. An automated early warning system, communicating warnings of varying urgency levels to ATCOs via a human-computer interface, is therefore indispensable [104]. 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 [105]. 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 [106]. 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 [22]. Under the situation of excessive warnings, operators might overlook or deprioritize the unexpected warnings, leading to a heightened risk of DF to the warnings [107].

Given the severe implications of DF, considerable attention has been paid to DF recognition, intervention design, and reduction [21, 24, 108]. Recognizing DF is a crucial initial step, serving as a reactive strategy to alert ATCOs about ignored or misunderstood warnings. A cognitive strategy framework for DF detection, comprising awareness-based and planning-based detection mechanisms, has been proposed in the literature [109]. Nevertheless, traditional DF identification methods, like self-reporting and retrospective analysis, are subjectively limited, as noted by [11]. Recent studies have shifted to objectively assessing physiological aspects such as eye movements and brain activities during monitoring [17, 75, 110]. 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 [111, 112]. 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 chapter 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 [113]. Vision is a vital channel for human-computer interactions in cognitive tasks, especially in supervisory functions. For example, [2] discussed the pivotal role of visual awareness in human-computer interface design, highlighting how the limitations in our perception can lead to significant DFs. Besides, [25] demonstrated the notable eye movement patterns linked to the degraded states of engagement. [114] also confirmed that the relations between saccade peak velocity and the first view time are affected by cognitive states.

3.1.2 Challenges

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) [30, 31]. 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 [19]. 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 [7]. Misinterpretation (MI) condition involves processing but misunderstanding warnings, causing erroneous foresight of warnings and thus necessitating enhanced training [21]. 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 [19]. 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 DFs and routine visual processing [72]. Additionally, gaze patterns related to DF are subtle and vary across individuals, further complicating the identification of consistent indicators [115]. 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 [28, 76].

3.1.3 Objective

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 [21] 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 [28]. In this phase, we test the following hypotheses: H3: the gaze features and warning frequency can be used to recognize four detection types.

This chapter 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.

3.2 Methodology

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 3.1, 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.

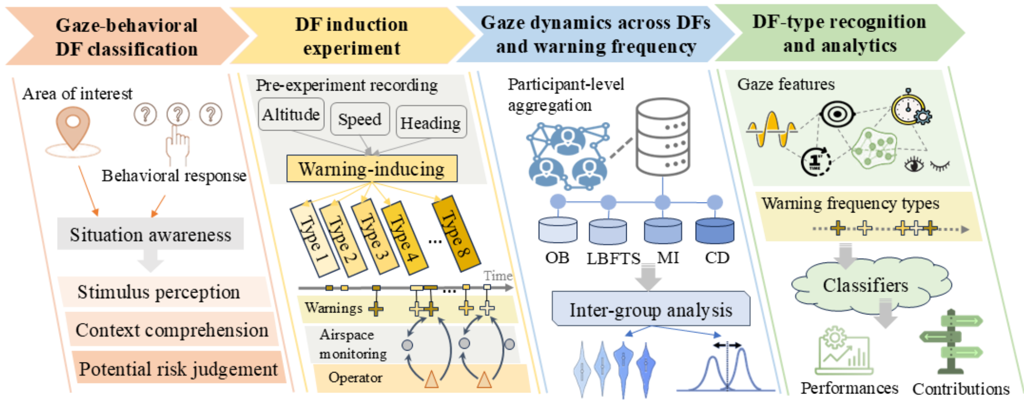


Figure 3.1: Flowchart illustrating the four-phase framework for DF classification and recognition.

3.2.1 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 [30]. 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 [116]. As shown in Figure 3.2, 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

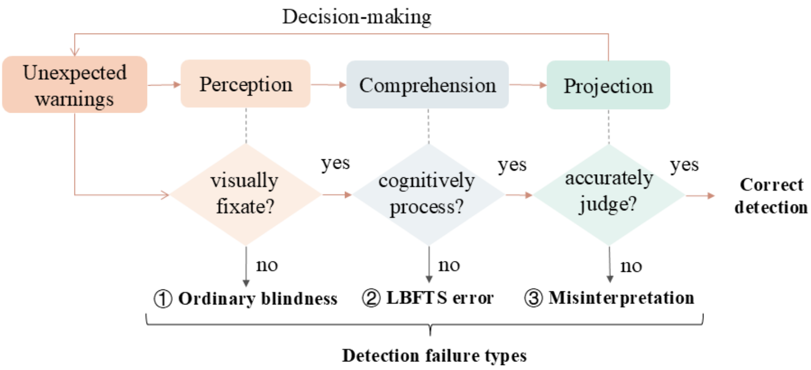


Figure 3.2: The proposed gaze-behavioral DF classification based on Endsley's situation awareness theory.

fixate on the warnings, are categorized as ordinary blindness (OB) to warnings [117]. 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 [7]. 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 [118]. 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 [119], 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.

3.2.2 DF induction experiment

3.2.2.1 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 [21]. All participants had either normal or corrected-to-normal vision. Informed consent was obtained from all participants prior to the commencement of the experiment, as shown in Appendix 6.2 and 6.2.

3.2.2.2 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 3.3. 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.

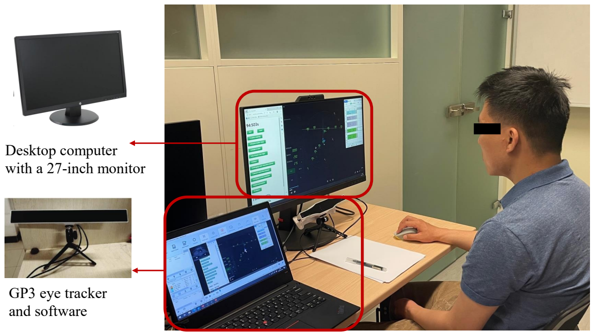


Figure 3.3: Experiment setting.

3.2.2.3 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 3.4.

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 ATC simulator designed to provide a user-friendly environment for simulating airspace management tasks, as de-

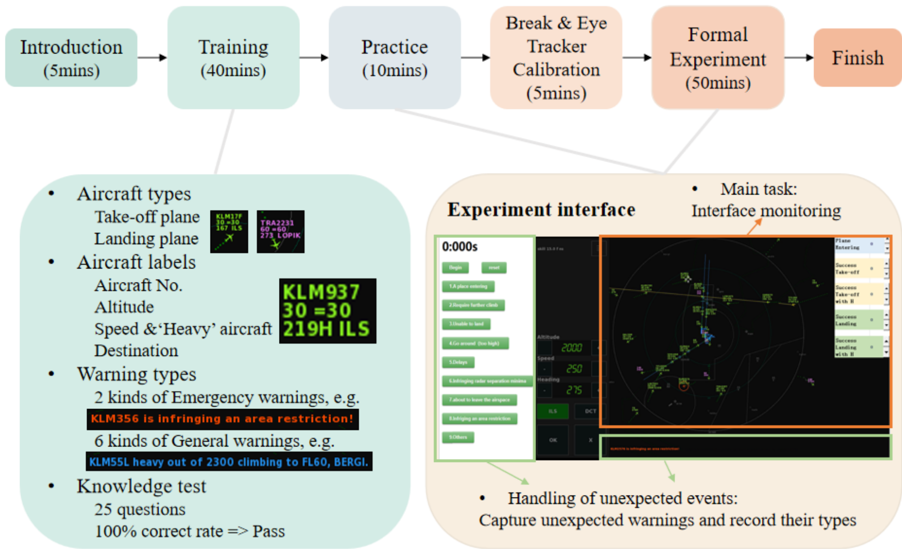


Figure 3.4: Experiment procedures.

scribed in [120]. 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 DFs 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 (see Table 6.1) 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 [121]. Detailed information about the tasks is shown below.

Airspace monitoring task. As shown in Figure 3.4, 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 ATC 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 3.1, 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 6.1 in Appendix I 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 [22], 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. This duration was thus adopted as the fixed analysis window, ensuring it captured the complete behavioral response cycle while providing a standardized timeframe for comparing time-sensitive eye-movement metrics (e.g., the first view time) across all trials. Each participant was exposed to 189 warnings over the 50-minute duration, reflecting the high frequency of alerts seen in real-world ATC environments [19] and ensuring sufficient data collection on DFs, which are rare occurrences.

3.2.2.4 Warning frequency settings

Two types of warning frequency are set in the experiment, including continuous warnings and interval warnings, as shown in Figure 3.5. 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.

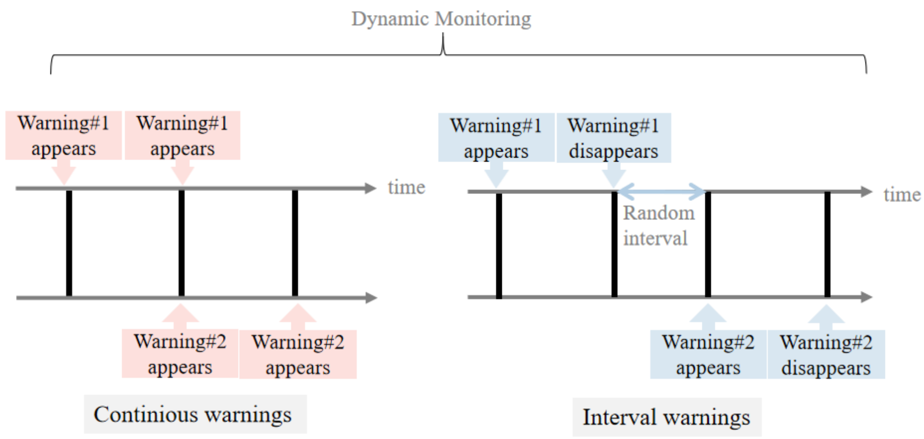


Figure 3.5: The settings of two types of warning frequency.

3.2.3 Gaze data collection

In the experiment, DFs were categorized based on subjects' gaze and response to warnings, as shown in Figure 3.6. 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 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

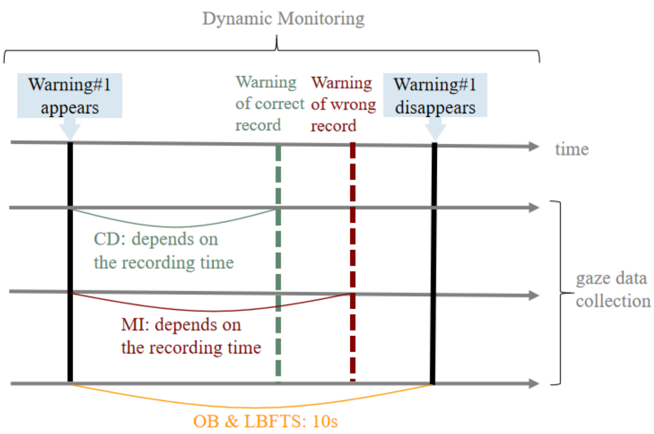


Figure 3.6: The eye movement data collection of different DF types during the experiment.

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 response.

3.2.4 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.

3.2.5 DF-type recognition and analytics

The process of DF-type recognition and analytics is shown in Figure 3.7. Two DF recognition tasks were compared: one with only eye movement inputs and another

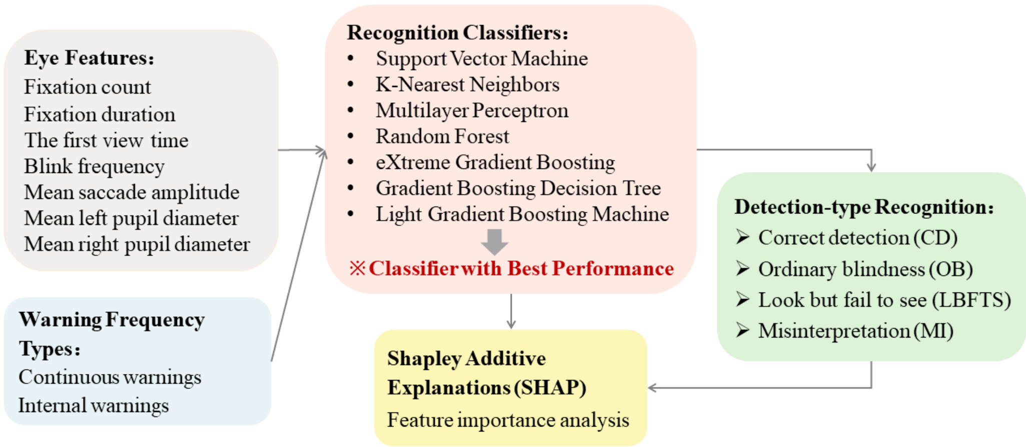


Figure 3.7: The process of DF-type recognition and analytics.

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 3.1, 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 [25, 122]. 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 [123]. 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 [79]. Light Gradient Boosting Machine (LightGBM), developed by Microsoft in 2017, stands out for its efficiency and performance in large dataset contexts [124]. 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 [120]. The output of each model is the classification of four detection types (i.e., OB, LBFTS, MI, and CD).

Table 3.1: 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 (MSA)	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 SHAP method, which quantifies each feature's mean marginal contribution over all feature combinations [25, 28]. 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:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_S(x_{S \cup \{i\}}) - f_S(x_S)] \tag{3.1}$$

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:

$$f(x) = \phi_0 + \sum_{i=1}^{|F|} \phi_i \tag{3.2}$$

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.

3.3 Results

3.3.1 Gaze dynamics across various DFs and warning frequencies

This subsection explores the gaze dynamics across diverse DF types. It also examines how different warning frequencies affect DF occurrence, as reflected by eye movement metrics, shedding light on their impact on attentional allocation.

3.3.1.1 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 3.2. 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 3.2 and allowed for an in-depth examination of each detection type's distinct eye-tracking features.

While the overall pattern of results provides strong support for H1, a detailed examination of individual metrics reveals one finding that warrants specific discussion: the asymmetry observed in pupil diameter measures. Specifically, the analysis showed a statistically significant effect for mean pupil diameter of right eye (MRPD) but not for its counterpart, mean pupil diameter of left eye (MLPD). This inter-ocular discrepancy can be understood through a layered consideration of physiological and methodological factors. First, physiological anisocoria—a naturally occurring, minor baseline asymmetry in pupil size between the two eyes—is common in the general population [125]. Since pupillometric analyses typically examine changes relative to a baseline period, pre-existing baseline differences can affect the sensitivity to detect task-evoked changes in each eye independently. In addition, the inherent measurement characteristics of eye-tracking systems introduce another layer of variability. Factors such as blinks, partial eyelid occlusion, or subtle calibration drift can differentially impact data quality between eyes, contributing to measurement noise [126]. Taken together, these factors provide a parsimonious explanation for the observed statistical discrepancy, attributing it to expected sources of variance rather than to a meaningful lateralized cognitive effect specific to warning detection. Crucially, this nuance does not compromise the validity of the primary findings concerning gaze patterns (FC, FD, MSA, FVT), which are derived from more robust, binocularly integrated or conjugate eye movement signals. However, the observation offers a valuable methodological insight for future research, suggesting the potential utility of employing binocular-averaged pupil metrics or implementing more stringent, eye-specific data quality criteria to enhance measurement reliability.

To further elaborate on the processed data and corresponding results, Figure

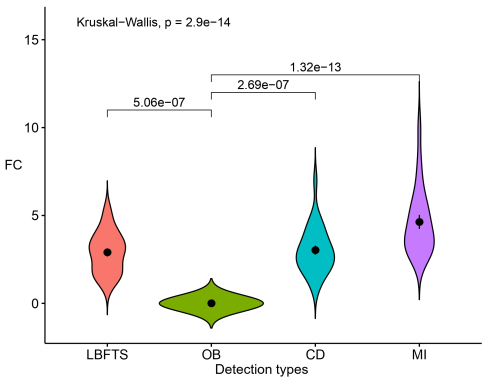
Table 3.2: Differential eye movement metrics across warning detection types.

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

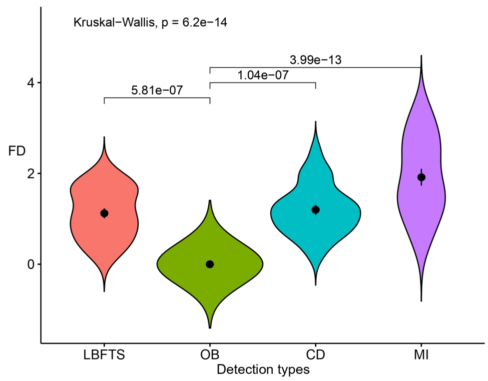
1, 2, 3, and 4 denote OB, LBFTS, MI, and CD, respectively.
* p-value less than .05.
** p-value less than .01.

3.8 presents violin plots of five metrics with significant differences, as indicated in Table 3.2. These visualizations are closely tied to the data processing techniques described in Section 3.2.4, where samples were consolidated on a per-participant basis and averaged for each individual before undergoing statistical testing. The shapes of each violin plot in Figure 3.8 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.

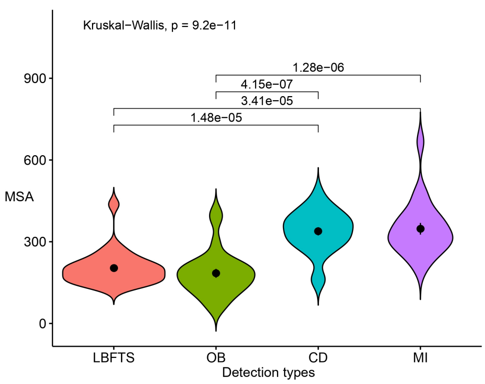
For both FC and FD metrics (see Figure 3.8a and Figure 3.8b), the OB detection type is markedly distinct, with significantly lower values indicating minimal or no fixations on warnings. This contrasts with the LBFTS and CD conditions, which display comparable distributions, suggesting a moderate level of attention to the



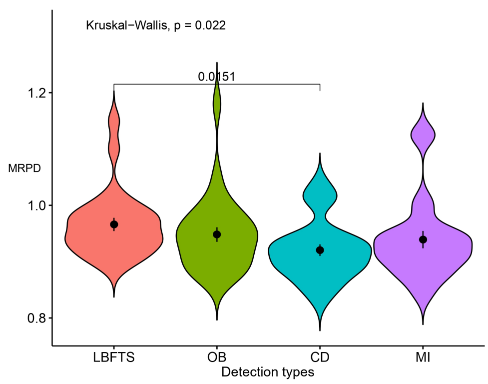
(a) FC



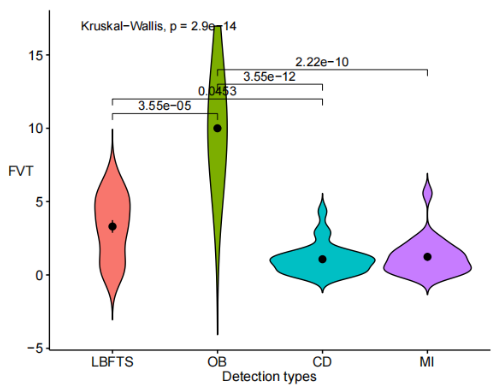
(b) FD



(c) MSA



(d) MRPD



(e) FVT

Figure 3.8: The violin plots of five metrics with significant differences across the detection types.

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 3.8c), 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 3.8d, the MRPD plot illustrates varying distributions of pupil dilation across detection types, with LBFTS and CD showing a statistically 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 3.8e, 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.

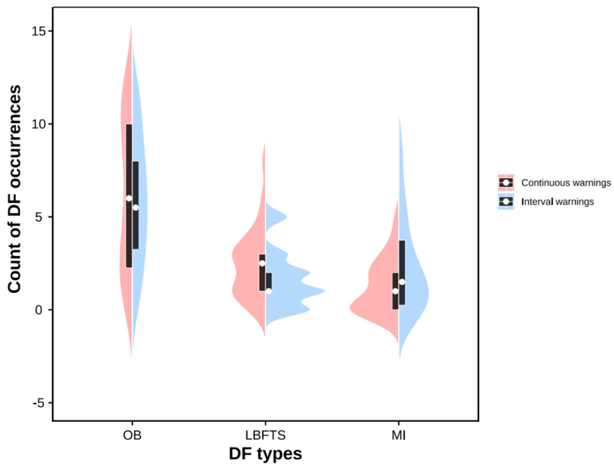


Figure 3.9: The split violin plot for comparative distribution of DF types across warning frequency conditions.

3.3.1.2 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 3.9. 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 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

Table 3.3: p-values of Mann-Whitney U test for continuous and interval warnings across three DF types and various eye-tracking features.

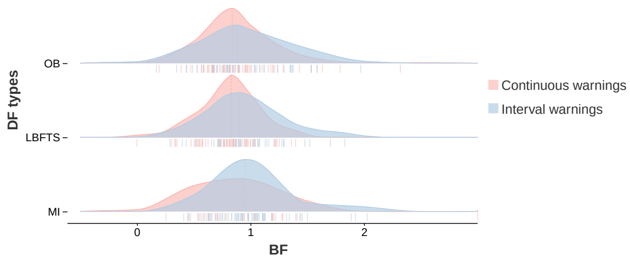
Eye-tracking features	p-values of Mann-Whitney U test for continuous and interval warnings across three DF types		
	OB	LBFTS	MI
FC	1.000	0.462	0.415
FD	1.000	0.980	0.124
BF	0.046*	0.038*	0.312
MSA	0.744	0.037*	0.048*
MLPD	0.398	0.314	0.487
MRPD	0.402	0.436	0.863
FVT	1.000	0.643	0.010**

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

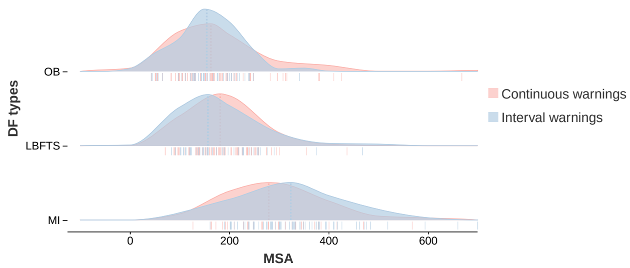
a relative insensitivity to continuous warning presence.

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 3.3, 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 3.10 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.

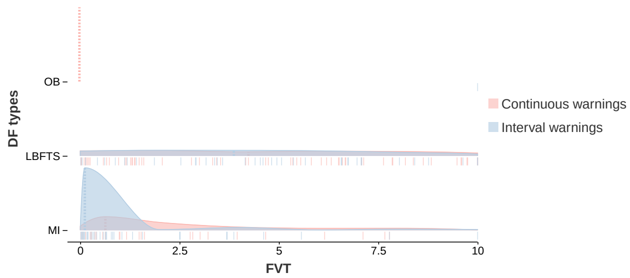
Figure 3.10a presents a ridgeline plot of BF differences among DF types (OB, LBFTS, and MI). The median lines, reveal a distinct BF distribution for OB and LBFTS under varying warning frequencies. Under continuous warnings, subjects



(a) Blink frequency



(b) Mean saccade amplitude



(c) The first view time

Figure 3.10: The ridgeline plots of three metrics with significant differences between continuous and interval warnings under different DF types.

show reduced and more uniform BF, implying potential 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 3.10b 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 3.10c, 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.

3.3.2 DF-type recognition and analytics

This section focuses on gaze-based DF-type recognition, which leverages eye movement data to distinguish DF types, and feature importance analysis by SHAP, which assesses the contribution of key features to the recognition process.

3.3.2.1 Gaze-based DF-type recognition

Table 3.4 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, 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 3.5, 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 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

Table 3.4: Performance of the compared models for classifying eye-tracking features into four detection types.

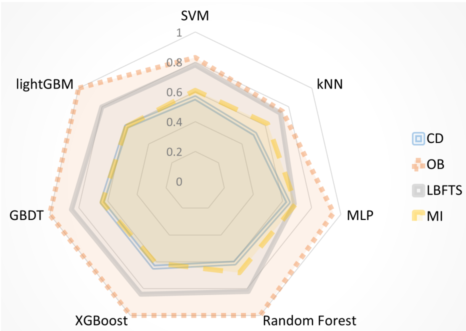
	SVM	kNN	MLP	RF	XGBoost	GBDT	LightGBM
Accuracy	0.67	0.65	0.69	**0.74**	0.67	**0.74**	0.73
Precision	0.68	0.65	0.69	**0.74**	0.66	**0.74**	0.73
Recall	0.67	0.66	0.69	**0.74**	0.67	**0.74**	0.73
F1 score	0.66	0.64	0.67	0.73	0.66	**0.74**	0.73

Table 3.5: Performance of the compared models for classifying eye-tracking features and warning frequency types into four detection types.

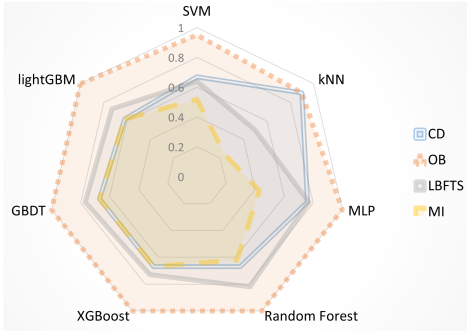
	SVM	kNN	MLP	RF	XGBoost	GBDT	LightGBM
Accuracy	0.69	0.63	0.68	**0.79**	0.76	0.77	0.74
Precision	0.69	0.66	0.67	**0.80**	0.77	0.78	0.75
Recall	0.69	0.64	0.68	**0.79**	0.77	0.78	0.74
F1 score	0.69	0.60	0.66	**0.79**	0.77	0.78	0.74

results of Table 3.5, as shown in Figure 3.11. Figure 3.11 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.

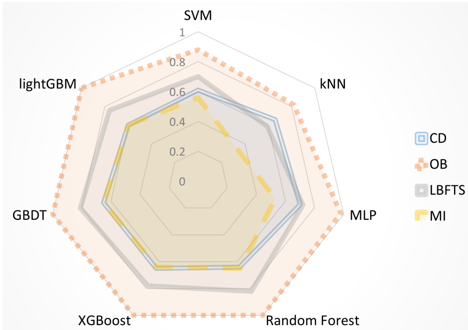
Within the OB type, defined by the lack of fixation on warnings, the models, particularly the four advanced tree-based models, exhibit outstanding accuracy, achieving precision and recall scores of up to 1.00. This exceptional performance highlights the unique gaze characteristics associated with OB, such as absent fixations and extended first view time as clarified by prior research [127], providing distinct indicators for classification.



(a) Precision



(b) Recall

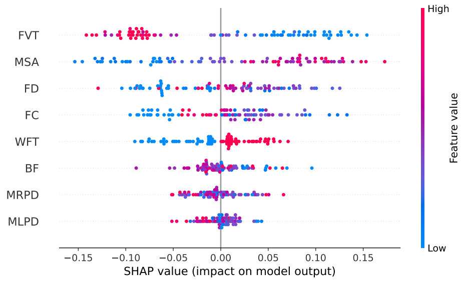


(c) F1 score

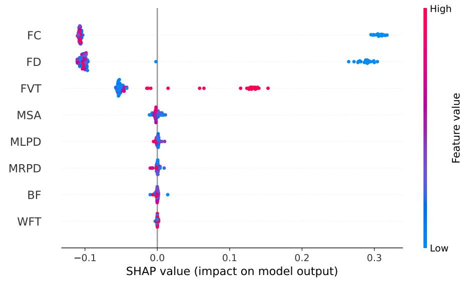
Figure 3.11: A comparative analysis of model performances from a single specific detection category.

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 precision 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. 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.

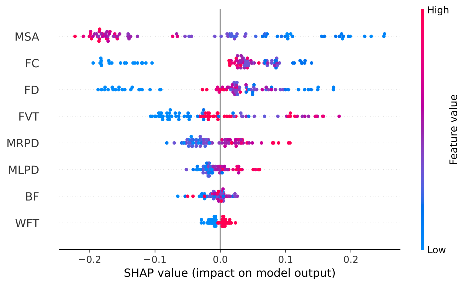
The MI and CD categories, representing incorrect and correct warning record-ings, respectively, exhibit distinct predictive challenges. MI slightly surpasses in precision, with MLP scoring 0.69, compared to CD's highest of 0.64. In contrast, CD excels in the recall, with kNN achieving 0.9, outdoing MI's 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.



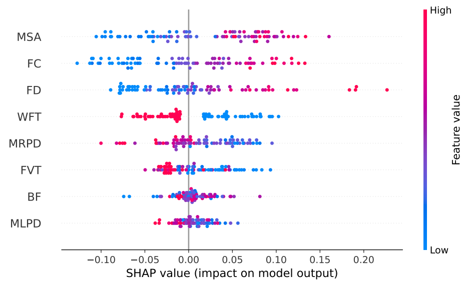
(a) Correct detection (CD)



(b) Ordinary blindness (OB)



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



(d) Misinterpretation (MI)

Figure 3.12: Beeswarm plot of importance ranking of the eye-tracking and warning frequency type (WFT) features.

3.3.2.2 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 3.12 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 subfigure (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 DF types, followed by FD and WFT. Specifically, in subfigure 3.12b, 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 3.12a, CD's quicker response time is captured by FVT as a critical feature. Moreover, for LBFTS and MI, as shown in subfigure 3.12c and 3.12d, 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 3.12a and 3.12d, two warning frequency types affect the recognition 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 [111].

3.4 Discussions

3.4.1 Gaze dynamics and importance analytics in DF-type recognition

The statistical results in Section 3.3.1.1 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 3.3.1.1. These metrics collectively help to identify OB and summarize the lack of visual and cognitive engagement with critical warnings, consistent with prior research [127]. 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 [128]. 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 [129]. 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 [130]. As highlighted in [67], 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 [131]. 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 [114], 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 [117]. While the statistical analysis in Section 3.3.1.1 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 [132]. On the other hand, MI is primarily distinguished by the scanning range, indicating a potentially less focused visual search. According to [133], 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.

3.4.2 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 3.3.1.2, 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 [14]. 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 [129].

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 Section 3.3.1.2 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 [111]. 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.

3.5 Concluding remarks

This chapter 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 GBDT 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 step offers valuable theoretical and practical insights into advancing DF recognition and intervention in complex environments.

Chapter 4

ATC-specific hierarchical DF recognition framework

In this chapter, an ATC-specific hierarchical framework is developed to improve recognition of visual DF through a novel eye-tracking feature set and LLM-driven evaluation. The hierarchical framework includes adaptive alert detection, GPT-4V-based attention assessment, and fixation coordinate-aware feature extraction to address DF complexity in realistic settings. The chapter is structured as follows: Section 4.1 presents the research motivation and challenges. Section 4.2 details the experimental setup. Section 4.3 explains the methodology, including alert detection, visual attention evaluation, and multi-domain feature extraction. Section 4.4 reports results on alert detection, attention assessment, and recognition performance across feature domains. Sections 4.5 and 4.6 provide discussion and concluding remarks.

4.1 Background

4.1.1 Motivation

When identifying visual DF, many studies rely on target fixations to measure whether an individual has detected the target [129]. For instance, when examining visual field defects, IB—the failure to detect unexpected but salient stimuli in the visual field—is the most frequently observed visual DF [10, 129]. The conventional criterion for IB is the absence of gaze on the stimulus [118, 134]. However, not fixating on a target does not necessarily mean it was not noticed [135]. This approach overlooks more complex scenarios, such as "look-but-fail-to-see" errors and "seeing-without-looking" phenomena—cases that reflect distinct attentional mechanisms. "Look-but-fail-to-see" errors occur when an individual fixates on a stimulus but fails to cognitively process its existence, resulting in no response [136]. This can be attributed to over-attention to a specific cognitive task, leading to a lack of situational awareness of unexpected stimuli despite direct visual engagement [30, 116]. On the other hand, "seeing-without-looking" describes a situation where a stimulus is detected despite the absence of direct fixation, primarily driven by covert attention and peripheral vision processing [33, 137]. This phenomenon is known as pre-attentive or automatic detection, where peripheral vision facilitates the unconscious processing of unexpected stimuli and triggers reflexive responses [44, 138]. Neurophysiological evidence further supports this mechanism, revealing distinct brain responses when peripheral stimuli are detected without direct foveal fixation [135]. Considering the above two complex scenarios would provide a deeper understanding of DF in visual tasks, particularly in high-stakes operational environments such as ATC.

Therefore, this study aims to propose a novel peripheral vision-aware visual detection classification to recognize involved DF. This classification encompasses two aspects: the gaze status on the stimulus and the individual's corresponding behavioral response. Specifically, when the gaze is directed at a target, a lack of response in this context suggests a "look-but-fail-to-see" error [21], while a corresponding response denotes "effective detection" [35]. Conversely, if the gaze is not directed at the target but a correct response occurs, it implies "pre-attentive detection" via peripheral vision [135]. Failure to respond without gaze indicates "peripheral vision neglect." Despite this categorization, the understanding of the visual properties underlying these four detection types is limited, and no empirical data are available to support these distinctions and accurately recognize the involved DF.

4.1.2 Challenges

Accurately recognizing different visual detection types via peripheral vision-aware visual detection classification in an ATC environment presents significant challenges. On the one hand, a major challenge is the inherent spatiotemporal uncertainty of alerts in dynamic natural scenes. This uncertainty complicates the detection of alerts and the verification of whether a controller's gaze has landed on them. For instance, controllers must monitor various dynamic displays, such as radar and auxiliary displays, for weather and airport information [21]. On the primary radar display, each aircraft is represented by a blip and a label containing information like callsign, flight level, and destination code, with short-term conflict alerts causing the label to turn red [3]. The spatiotemporal uncertainty of alert positions and timings makes it impractical to use traditional pre-defined Areas of Interest (AOIs) to determine if a controller's gaze has landed on an alert [75]. Thus, eye-tracking object detection approaches using YOLO (You Only Look Once) were recently proposed in [139] and [140], but they have limitations. While the YOLO system is pre-trained to recognize common objects such as vehicles and animals, the ATC radar interface is unique and not composed of typical physical entities, requiring a more tailored alert recognition method [3]. Additionally, verifying whether a controller's gaze has landed on the alert after alert detection remains time-consuming and prone to human error. In recent years, the rise of generative AI and vision-language models like CLIP offers promising solutions for overcoming challenges in complex scenarios, such as task-based zero-shot transfer for distracted driver activity recognition, highlighting their potential in intelligent behavior evaluation [96, 141]. Large Language Model (LLM) has been applied to automate annotation tasks, particularly through natural language understanding and semantic labeling, where human verification is combined with LLM efficiency to improve accuracy in complex datasets [142, 143]. Inspired by this, applying the image processing capabilities of LLMs to assess human visual attention on alerts offers a novel approach to enhancing alert detection processes [140, 144].

On the other hand, whether gaze trajectory can adequately reflect peripheral vision processing has not been explored [145]. Traditional eye-tracking features like blink rate, pupil size, and fixation duration, while extensively used to assess visual attention in the central visual field [26, 28, 113], tend to oversimplify the complexities of human visual processing, especially regarding peripheral awareness [146]. Peripheral vision plays a critical role in detecting stimuli outside the direct line of sight, relying on broader visual tracking mechanisms that require different visual strategies, such as scanning movements and subconscious tracking, which are not reflected in traditional eye-tracking features [25, 44, 137]. These broader patterns may involve rhythmic scanning and temporal shifts critical for detecting peripheral stimuli. As indicated in [147], features integrating graph theory provide a more precise interpretation of the visual tracking dynamics. Although meeting the prerequisite of a graph neural network is challenging due to non-coinciding fixations in visual trajectories, its concept inspires us to fully explore the multi-layer information of visual trajectories. For instance, by employing spatial analysis, Fourier transform, and time-variant autocorrelation—methods that span multiple levels of signal processing, from spatial distribution to frequency analysis and temporal dynamics—we can uncover novel multidimensional embedding features [83, 148].

To address the aforementioned challenges, this paper proposes a hierarchical peripheral vision-aware detection failure recognition framework to assess human attentional states toward ATC-specific alerts. By automating the recognition of DF, the proposed framework can be integrated into air traffic management systems to provide effective feedback on controller attention states. This capability not only improves the reliability of alert detection in ATC, but also establishes a foundation for the development of an intelligent, human-centered ATC support system.

4.2 Experimental setup

4.2.1 Participants

This study evaluated the hierarchical framework via an ATC simulation at the Hong Kong Polytechnic University with 38 participants (age M = 23 yr, SD = 13.60, Max = 38 yr, Min = 19 yr), including 27 pre-professionals with ATC qualification certificates and 11 members from VATSIM (Virtual Air Traffic Simulation Network) Hong Kong, a global online community for realistic ATC simulation, all with over 1 year of simulated ATC training and simulated ATC experience, ensuring ATC proficiency and consistency. All participants had normal or corrected-to-normal vision. To ensure data security, all visual data in ATC simulation experiment were anonymized, with only relevant segments (e.g., eye movement trajectories and specific alert regions) processed by GPT-4V. Participants provided informed consent, and the study was approved by the IRB (HSEARS20211117002), adhering to ethical standards and guidelines.

4.2.2 Apparatus

In this study's methodological framework, a desktop computer with a 27-inch monitor (1920*1080 pixels) was paired with a Gazepoint 3 (GP3) eye-tracker to collect detailed eye movement data, as shown in Fig. 4.1. The eye-tracker software operates on a separate laptop connected to the monitor via HDMI. The GP3, an integrated desktop eye-tracking system, records various eye movement metrics at a sampling rate of 60 Hz, using binocular tracking within an operating distance of 50 cm to 80 cm from the monitor. The system's spatial accuracy ranges from

0.5° to 1° of visual angle, with a spatial resolution of 0.1° RMS. Calibration was performed using a 9-point method, with successful calibration requiring accurate fixation on each point. Calibration was checked regularly, and recalibration was performed if discrepancies were detected. Additionally, EEG data was collected using a synchronized headset to support future gaze-neural correlation analyses in ATC DF, with pre-experiment comfort assessments confirming no significant discomfort and ensuring data quality[54, 149].

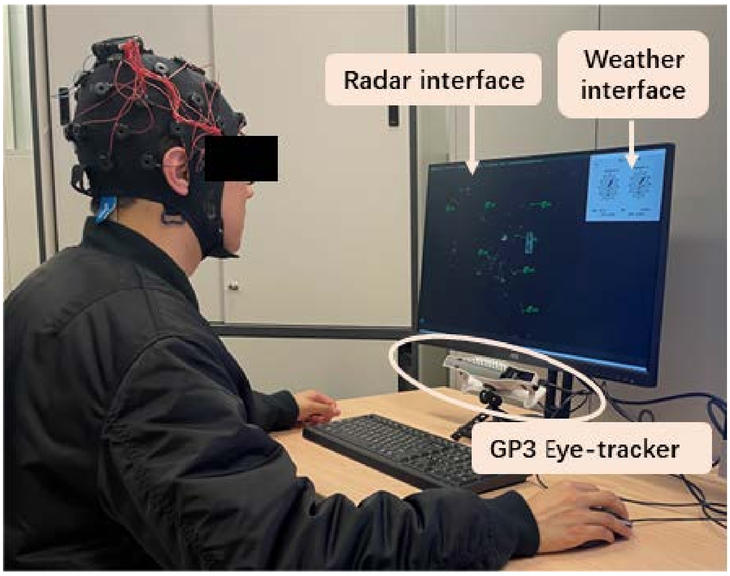


Figure 4.1: Experimental setup showing a 27-inch monitor (1920×1080 pixels) paired with a Gazepoint 3 eye-tracker, recording eye movements data.

4.2.3 Experiment procedures

The experimental procedure was designed to simulate the complex environment characteristic of ATC operations. The experimental interface, depicted in Fig. 4.2, comprises a radar-map interface from the Euroscope platform and a simplified weather interface, as informed by prior work on ATC human factors[3]. Partici-

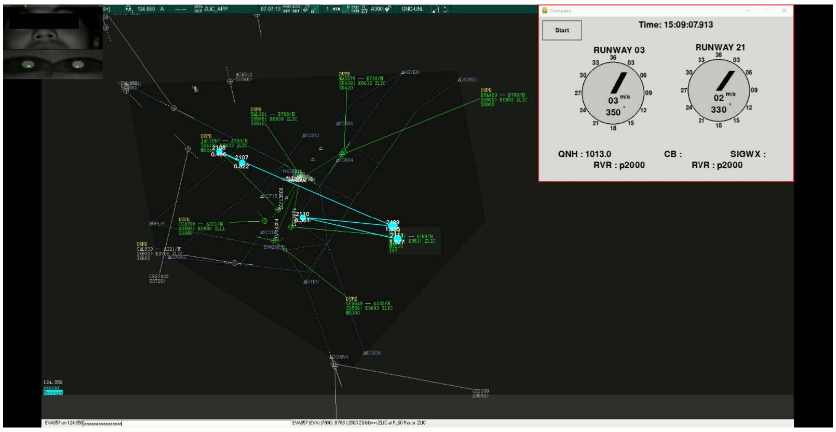


Figure 4.2: Experimental interface for the ATC multi-tasking scenario.

pants engaged in a 40-minute multi-tasking scenario, managing the safe arrival of 17 aircraft, each requiring specific descent and approach instructions. This scenario encompassed radar surveillance, issuing clearances, weather monitoring, and radiotelephony communication with a pre-arranged pseudo-pilot, an experienced ATC professional who simulated realistic flight scenarios to create an immersive operational environment[3]. Additionally, the scenario incorporated randomly appearing aircraft, termed fly-over flights, crossing the radar airspace—a common feature in real ATC operations. These flights, while not requiring direct participant intervention, increased cognitive load by necessitating sustained situational awareness in a busy airspace. Overall, the experimental duration was designed to simulate the sustained cognitive load experienced in practical ATC operations while ensuring sufficient data collection for robust analysis. It aligns with similar studies, such as [150], which employed around 40-minute simulated driving tasks to simulate high-stress environments and collect robust data.

During each 40-minute ATC simulation session, a minimum of 13 alerts per participant (M= 16, SD= 2) were randomly triggered, with additional alerts varying based on the simulation's evolving traffic conditions, to simulate comprehensive unpredictability of alert presentation and operational demands in multitasking conditions, aligning with previous studies (e.g., [151] used probabilistic alerts; [152] employed variable alert scenarios). The distribution of alerts across participants is detailed in Fig. 6.1 in Appendix II. Post-experiment feedback validated the design's workload appropriateness, with participants reporting a workload of 5–7 on a 1–9 scale, indicating a challenging yet manageable cognitive demand. Throughout each 40-minute session, a total of 13 pre-programmed alerts were randomly triggered to evaluate participants' ability to detect and respond to sudden visual stimuli amidst concurrent tasks. Each alert, defined as a single data point, lasted 5 seconds and appeared individually at random intervals and locations. These alerts were initiated by the pseudo-pilot, who randomly selected one fly-over flight to simulate an emergency by activating the 7700-emergency squawk code—an ATC standard for distress signals[153]. These brief, attention-grabbing signals, illustrated in Fig. 4.3, caused the label of an aircraft to turn red, indicating potential conflicts. This approach ensured that the timing and location of each alert were unpredictable, reflecting the dynamic and complex nature of real-world ATC operations. Participants were instructed to monitor these alerts peripherally while focusing on their primary tasks. Upon noticing an alert, they were required to press the "A" key, providing a measurable indicator of their attentional engagement. This setup ensured a dynamic assessment of attention allocation under realistic, unpredictable conditions. Additionally, based on the current control situation, they could engage with the pseudo-pilot of the alerted aircraft, inquiring about flight intentions to assess priority levels and ensure air traffic safety. The alerts automatically resolved after 5 seconds, preventing additional cognitive load from interfering with their ongoing multi-tasking responsibilities.

In our experiment, we focused on the radar display as the primary information source due to its critical role in ATCOs' workflow. Radar alerts, such as those triggered by aircraft proximity violations, are the most immediate and high-priority tasks for controllers[58, 154]. We examined responses to single radar alerts, a rare but critical aspect of ATC operations[155]. This aligns with prior research on isolating critical events to identify distinct visual and behavioral patterns[58, 153], aiming to establish a foundational understanding of DF and their eye-tracking characteristics.

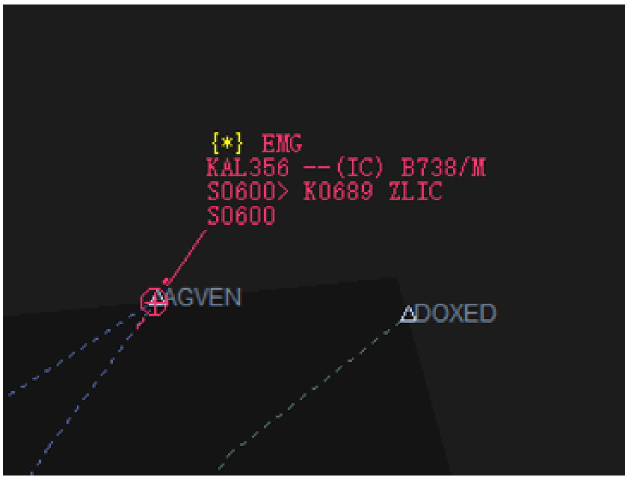


Figure 4.3: Example of a pre-programmed alert on the radar interface, showing a red aircraft label indicating a potential conflict.

4.2.4 Gaze data collection

In this study, gaze data were collected within a 7-second window following the onset of each pre-programmed alert, comprising a 5-second alert presentation duration and a 2-second post-alert response buffer. This 5-second alert duration ensures consistent reaction times [156], while also optimizing response effectiveness in emergencies as noted by [36]. However, in multitasking scenarios, the process from perceiving an alert to responding often involves attentional allocation or task-switching demands, resulting in a temporal lag, which necessitates the 2-second response buffer to capture delayed responses. This buffer, further supported by [22], accounts for residual attention processing, as controllers may require additional time to process and respond. As a result, the 7-second window comprehensively captures the complete gaze-behavior sequence, enabling accurate detection and response assessment in alignment with our framework for classifying DF. The timeline of this process is illustrated in Fig. 4.4, which depicts the 7-second window, including the alert presentation duration, response buffer, and gaze-behavior sequence, while highlighting potential delayed responses. Consequently, each 7-second interval is treated as an individual sample for detailed behavioral analysis.

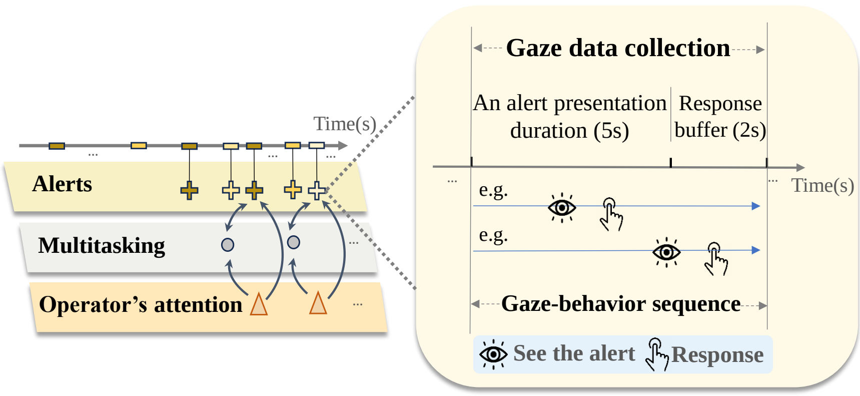


Figure 4.4: Timeline of the 7-second gaze data collection window, comprising a 5-second alert presentation duration and a 2-second response buffer.

4.3 Methodology

This study first classifies visual detection types by integrating peripheral vision tracking and human attentional states derived from human fixations and behavior. A **Hierarchical Peripheral Vision-aware Detection Failures Recognition Framework** specific to ATC alerts is further developed to recognize the proposed DF types, as depicted in Fig. 4.5, and validated through an ATC simulation experiment yielding empirical data. The framework comprises three phases: (1) an Adaptive Symbolic Alert Detection method is introduced to identify alert regions with spatiotemporal uncertainty on the ATC radar interface; (2) a GPT-4V-powered evaluation of visual attention engagement to alerts is employed to facilitate labeling samples into four detection types; and (3) a Fixation Coordinate-Sensitive Multi-Domain Feature Set is proposed to analyze fixation dynamics to classify the four detection types, comparing domain-specific and integrated feature performance.

4.3.1 Peripheral vision-aware visual detection classification

The detection types in this study are categorized based on a combination of gaze behavior and participant responses, as illustrated in Fig. 4.6. The figure outlines a decision-making process starting with an alert's appearance and branching into different outcomes based on whether the participant's gaze was directed at the alert and their timely response. These detection outcomes serve as the ground truth for evaluating detection methods, ensuring consistency and a well-defined standard for classification tasks.

Firstly, peripheral vision processing is characterized by the appearance of an

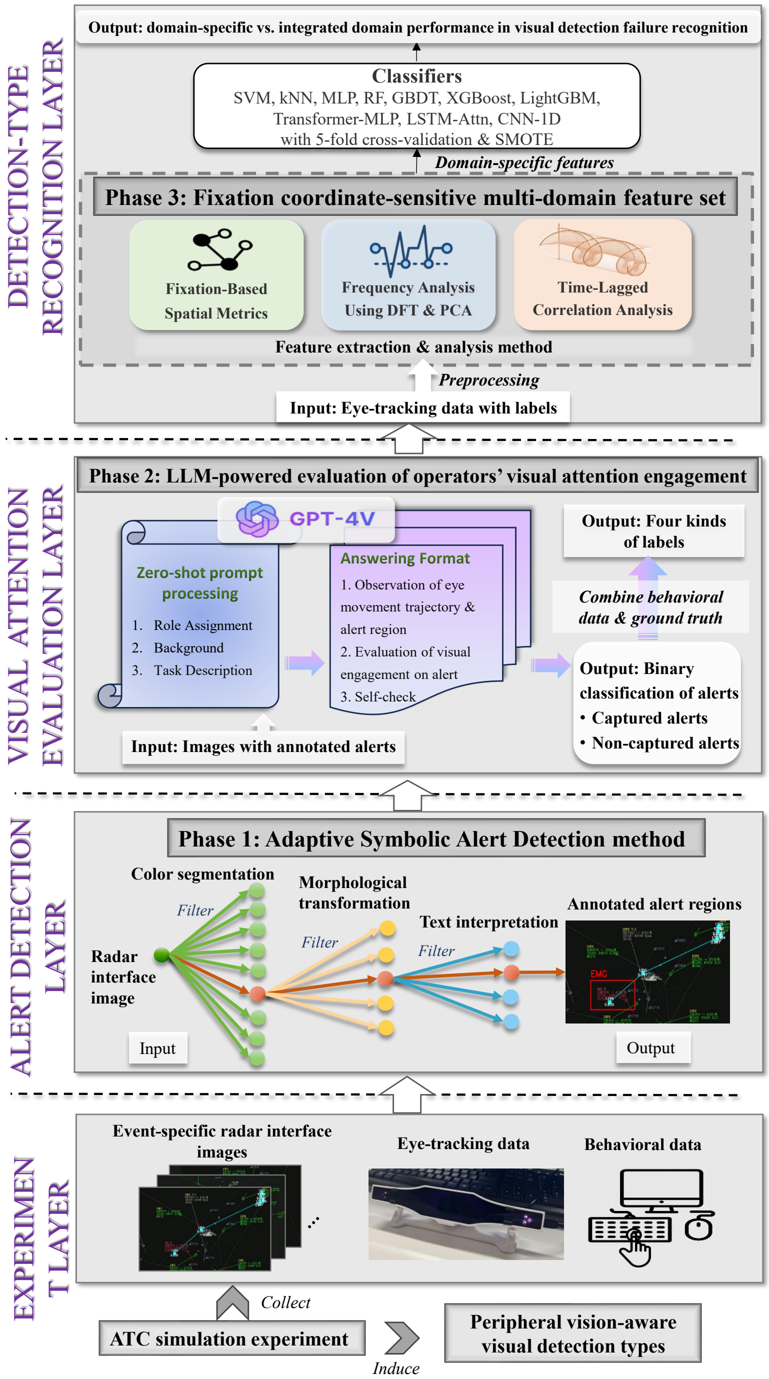


Figure 4.5: Hierarchical peripheral vision-aware DF recognition framework for ATC-specific system, integrating ATC simulation, alert detection and annotation, LLM-based visual attention evaluation, and multi-domain feature analysis.

alert within the participant's peripheral vision, a critical phase that determines whether the alert will be subsequently attended to [157]. If the participant responds to the alert without directly gazing at it, the incident is classified as Pre-Attentive Detection [44, 138]. This indicates that peripheral vision was effective in detecting the alert, even though the operator's focus remained on other tasks, possibly delaying their attention to the alert itself. This classification will be essential in evaluating detection methods that leverage peripheral awareness, as it establishes the baseline for identifying alerts that were recognized without direct gaze. In contrast, Effective Detection occurs when the participant not only shifts their gaze toward the alert but also responds within the allotted time [35]. This suggests that the participant visually processed and cognitively acknowledged the alert, considering its implications for the broader radar environment and aircraft management. The Effective Detection category provides a critical benchmark for assessing methods designed to ensure timely and appropriate responses to alerts, serving as a reference point for successful alert recognition and reaction. Another critical classification is the Look-but-Fail-to-See (LBFTS) Error[12, 136]. This occurs when the participant's gaze does land on the alert, yet they fail to respond within the designated time frame. Such an outcome indicates that while the alert was visually perceived, it was not cognitively processed in a manner that prompted an appropriate reaction [33]. This error highlights a breakdown in the link between visual attention and subsequent cognitive action. By including LBFTS Errors in our classification, we provide crucial ground truth for identifying situations where visual perception does not translate into an effective response. This allows for the analysis of cognitive processing failures in detection tasks. Lastly, Peripheral Vision Neglect is observed when the participant neither gazes at the alert nor responds within the time limit, indicating a failure to detect alerts through peripheral vision. This type is essential to assess, as peripheral vision is crucial for detecting unexpected events and maintaining situational awareness without direct fixation, as highlighted in [33]. This classification will be used as ground truth to evaluate the worst-case scenarios in detection methods, where both visual attention and peripheral awareness fail, offering insights into the limitations and challenges in alert detection.

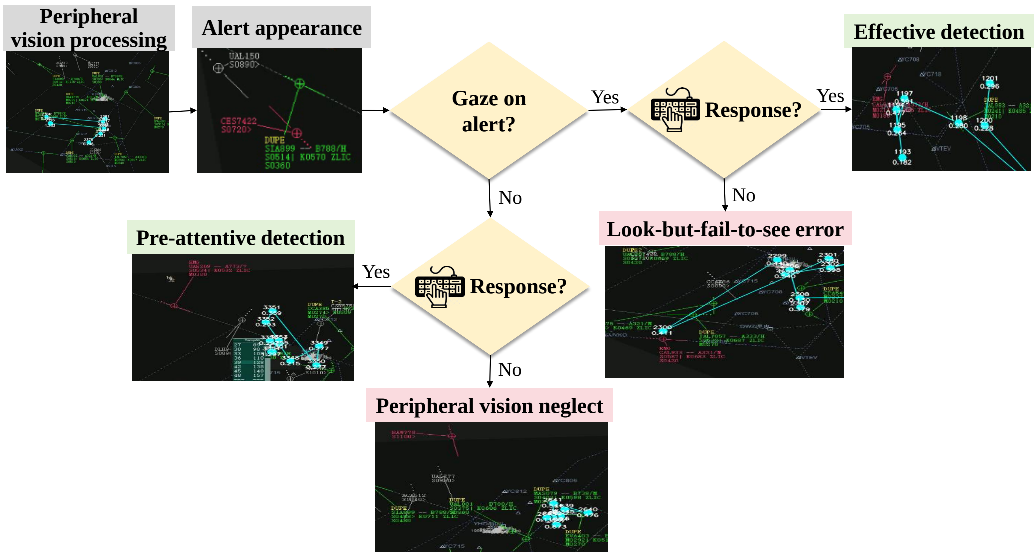


Figure 4.6: Decision tree classifying peripheral vision-aware visual detection types based on gaze and response behavior in ATC alert scenarios.

4.3.2 Phase 1: Adaptive symbolic alert detection method

This study implements a robust image processing framework designed to enhance the detection and annotation of critical alerts on ATC radar interfaces. Each step in our methodology is precisely calibrated to optimize accuracy and efficiency, drawing from advanced computational techniques to process images at a resolution of

512x288 pixels. The image processing was implemented using Python 3.11, with the following libraries employed: OpenCV (version 4.9.0.80) for image processing tasks such as color segmentation and morphological operations, and NumPy (version 1.23.5) for handling array operations and matrix manipulations. Phase 1, shown in Fig. 4.5, outlines the steps of the adaptive symbolic alert detection method. The term ”Adaptive Symbolic” reflects the method's ability to dynamically process and integrate various types of symbolic information—such as colors, contours, and text—each representing different aspects of an alert. By adapting to the specific characteristics of these symbolic elements, the approach can effectively detect and annotate alerts in a manner that is both contextually aware and highly precise, ensuring that the system responds appropriately to the complexity of the information it encounters.

The first stage involves effective color segmentation techniques to isolate alerts based on color cues, specifically red, commonly used in ATC systems to denote urgency. The process begins with converting the input images to the Hue, Saturation, Value (HSV) color space to better distinguish color variations under different lighting conditions [158]. Specific HSV ranges are targeted:

- Lower red range: HSV values of (0, 120, 70) to (10, 255, 255)
- Upper red range: HSV values of (170, 120, 70) to (180, 255, 255)

These thresholds are chosen based on empirical data indicating their effectiveness in capturing a wide spectrum of red tones displayed on diverse ATC interface technologies. Binary masks are then generated to identify areas of the image falling within these red ranges, using OpenCV's `cv2.inRange()` function.

The second stage, morphological transformations, refines the segmented regions to ensure accurate detection. This involves applying a series of filters:

- A closing operation with a 7x7 pixel kernel to close small holes and connect broken regions within the alert areas, ensuring continuity.
- An opening operation to remove minor noise points, followed by dilation to enhance the visibility and distinctness of the contours.

These transformations are crucial in preparing the segmented areas for precise contour detection, enabling the identification of critical elements such as circular symbols (representing aircraft icons), associated text labels (displaying flight information and status), and connecting lines. The processed image undergoes contour analysis to isolate these components, with the Hough Line Transformation employed to detect any linear elements connecting the symbols and labels.

The third stage, text interpretation, integrates all identified elements into a cohesive alert region. In this step, the detected contours and lines are merged into a single bounding rectangle that encompasses the entire alert, including both the circular symbols and the associated text labels. This unified region is then annotated with ”EMG” above it. The annotation serves as a clear and precise interpretation of the alert’s nature, ensuring that it is properly highlighted and ready for further analysis by LLMs. This comprehensive process effectively prepares the images to assess alert visibility and attention tracking in ATC settings.

This method effectively segments, refines, and integrates critical alert elements—such as aircraft icons and flight information—into a unified region. By accurately detecting and annotating these components, the system ensures clear and reliable identification of critical alerts, optimizing the images for further analysis in ATC settings.

4.3.3 Phase 2: GPT-4V-powered evaluation of visual attention engagement

This study tackles the challenge of manually verifying a controller's attention to alerts by utilizing GPT-4V, a visual extension of the GPT-4 model, in conjunction with LLMs. Phase 2, depicted in Fig. 4.5, outlines the refining summary of LLM-powered evaluation of operators' visual attention engagement on alerts. Specifically, the `GPT-4o` model was selected for its advanced capabilities in visual and language processing, deemed suitable for handling the complex visual inputs typical in ATC scenarios. The model provided a nuanced and context-aware analysis of controllers' visual attentional engagement with alerts, as detailed in Algorithm 1.

To generate initial binary labels, we employed zero-shot prompts tailored to the specific context of ATC interfaces[159]. Each event-specific image, annotated with alerts, was encoded into a suitable format and input into the GPT-4V model. The prompt structure defined the model's role, background context, and task objectives. It was instructed to act as a researcher analyzing eye movement trajectories in ATC tasks, specifically assessing whether blue trajectory markers (fixation trajectory) fell within red emergency zones (alert region). The task was to classify eye movement trajectories into Captured alert or Non-captured alert based on gaze engagement with the alert region. Non-captured alerts were assigned when no fixations were detected within the designated alert regions, while their presence indicated a Captured alert.

Algorithm 1 Evaluating visual attention engagement using GPT-4V

Require: Image dataset directory, OpenAI API key
Ensure: Responses are saved to an Excel file

```
Initialize variables:
    openai.api_key ← Retrieve from environment variable or set manually.
    folder_path ← Path to the directory containing images.
    image_paths ← List all image files in the specified directory.
    models ← ['gpt-4o'].
Define prompt for analyzing ATC images:
```

- Role Assignment:

```
Assume the role of a researcher analyzing eye movement trajectories in ATC
control tasks.
```

- Background:

```
Analyze images from a simulated ATC interface showing various detailed
information with ['status indicators'], ['flight information'], ['emergency
alerts'], ['emergency highlighting'], ['eye movement visualization'].
```

- Task Description:

```
Determine if eye movement trajectories focus on areas highlighted for
emergencies. Verify the presence of blue trajectory markers within
red emergency zones, classifying observations as 'Captured alert' or
'Non-captured alert' based on their alignment.
```

```
Process each image using the defined models:
for each image_path in image_paths do
    Convert image to base64 encoding
    for each model in models do
        Construct payload for API request
        Send POST request to OpenAI's API
        if response is successful then
            Parse and save response
        else
            Log error details
        end if
    end for
end for
Compile responses into a DataFrame
Save the DataFrame to an Excel file at the specified path
```

This classification approach considers the distinction between central and peripheral vision, a key factor in assessing gaze engagement. Central vision, defined by the fovea, spans approximately 2° of visual angle for high-acuity perception, while peripheral vision extends beyond this range to 90°–100° eccentricity, with decreasing detail sensitivity[73, 160]. Rather than applying a fixed 2° range, we assess central vision engagement by determining if the fixation trajectory overlaps the alert region. This approach is grounded in three primary considerations: its theoretical basis in foveal dynamics, its alignment with ATC task demands, and its methodological efficacy. First, fixations, recorded as stabilized gaze positions (200–300 ms), dynamically reflect foveal focus [160]. The fixation trajectory embeds the 2° foveal span within real-time gaze behavior, aligning with the fovea's natural targeting process without imposing a fixed angular limit. Second, since ATC alerts vary in size and position due to factors like aircraft labels and emergency trigger, the annotated alert regions are directly aligned with task-relevant targets, avoiding mismatches a fixed range might introduce [3]. Third, a 2° range demands pixel conversion and overlap assessment with a hypothetical circular zone per fixation point, increasing computational complexity and potential mismatch with variably sized alerts. Our approach relies on the direct spatial relationship between fixation trajectory and the alert region, providing a straightforward and efficient criterion.

To ensure the reliability of binary labels, the model sent payloads to the OpenAI API for parsing and analysis, with each image requiring approximately 750 tokens for processing and generating categorical labels based on prompt-based reasoning. This processing, involving batch analysis of eye movement trajectories collected post-data acquisition, results in delays typically ranging from 320 milliseconds to 1 second per request, ensuring data integrity. Moreover, to validate these labels, we aligned the LLM outputs with ground truth labels established through expert evaluations. Therefore, the final labels used for model classification were 100% consistent with the ground truth. This method is similar to that described in [80], where a natural prompt was used to generate initial LLM labels, which were then refined to align the outputs with ground truth labels. Alternative methods, such as traditional AOI settings and computer vision techniques, offer insights in some cases but have limitations in intelligently assessing visual attention dynamics in relation to alert regions in ATC scenes[161]. Thus, expert verification remains the most suitable comparison for this task. Finally, the binary labels (Captured/Non-captured alerts) were further refined by integrating behavioral response data to derive four visual detection types: Pre-attentive detection, Peripheral vision neglect, LBFTS error, and Effective detection. By integrating LLM outputs with ground truth verification, we minimized the need for extensive manual labeling while maintaining methodological rigor.

4.3.4 Phase 3: Fixation coordinate-sensitive multi-domain feature set

To comprehensively capture the nuanced patterns of visual tracking dynamics, we propose a novel fixation coordinate-sensitive multi-domain feature set. This approach integrates spatial, frequency, and time-based features derived from gaze and fixation coordinate data, leveraging the synergistic interplay among these domains to provide a robust representation of gaze behavior. As illustrated in Fig. 4.7, this section will detail how these dimensions are combined to form a compre-

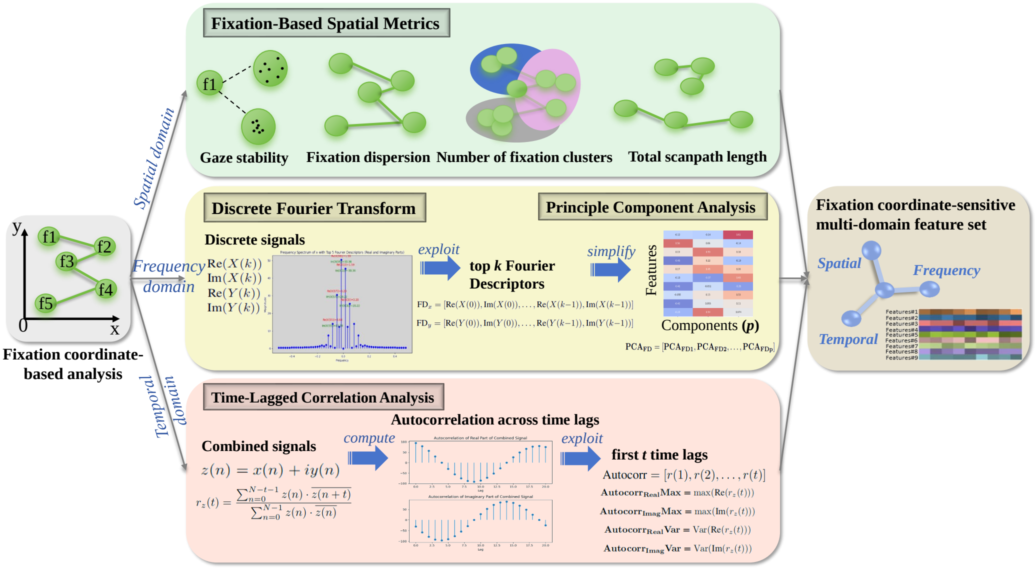


Figure 4.7: Overview of the proposed fixation coordinate-sensitive multi-domain feature set, combining spatial metrics, frequency descriptors, and time-lagged autocorrelation for analyzing fixation trajectories in ATC alert detection tasks.

hensive feature set for peripheral awareness detection analysis.

4.3.4.1 Fixation-based spatial metrics

The analysis of gaze behavior in this study incorporates fixation-based spatial metrics, which are essential for quantifying how visual attention is distributed across the visual field. These metrics include gaze stability, fixation dispersion, the number of fixation clusters, and total scanpath length, each contributing to a comprehensive understanding of visual attention dynamics.

Gaze stability is assessed by calculating the standard deviation of positions in pixels of gaze points that belong to each fixation within a specified time window [25]. Mathematically, if (x_i, y_i) represents the coordinates of gaze points within a fixation, the gaze stability σ_{gaze} is given by:

$$\sigma_{\text{gaze}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - \bar{x})^2 + (y_i - \bar{y})^2\right]} \tag{4.1}$$

where $\bar{x}$ and $\bar{y}$ are the mean coordinates of the gaze points in the fixation. This metric reflects the consistency of gaze within a fixation, with lower values indicating more stable gaze and higher values suggesting greater variability.

Fixation dispersion is evaluated by calculating the root mean square of the distances from each fixation point to the average fixation position within a time window [162]. This metric provides insight into how spread out the fixation points are relative to their mean position. The fixation dispersion $D_{\text{dispersion}}$ is calculated as:

$$D_{\text{dispersion}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}\right]^2} \tag{4.2}$$

This measure helps to quantify how widely the participant's attention is distributed during the task, with larger values indicating greater dispersion of fixation points.

The **number of fixation clusters** is determined using clustering algorithms such as k-means, which identify groups of fixation points that are spatially close. Each cluster represents a distinct area of interest that repeatedly attracts the participant's gaze. The number of clusters C indicates how many different regions of interest were focused on during the task. A higher number of clusters may reflect more distributed attention across multiple areas, whereas fewer clusters suggest more focused attention.

Finally, **total scanpath length** L is calculated to evaluate the overall distance traveled by the gaze across all fixation points [147]. It is a measure of the visual exploration strategy employed by the participant and is defined as:

$$L = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \tag{4.3}$$

A longer scanpath length suggests a more exhaustive search behavior, potentially indicating the need for more visual information gathering, while a shorter scanpath indicates a more efficient and focused scanning pattern.

Together, these fixation-based spatial metrics provide a detailed characterization of how participants allocate their visual attention across space, offering critical insights into the mechanisms that drive effective visual processing and peripheral awareness during task performance.

4.3.4.2 Frequency analysis using discrete fourier transform

In the frequency domain, we analyze the horizontal (x) and vertical (y) coordinates of fixations separately, as each direction may exhibit distinct frequency characteristics. Separating them ensures that we capture their individual oscillatory patterns and periodic behaviors without interference from cross-dimensional interactions. The frequency domain features derived from Fourier Descriptors (FDs) and Principal Component Analysis (PCA). These methods allow us to extract meaningful characteristics from the gaze data and distinguish between different detection types by focusing on the spatial distribution and movement patterns of fixation points.

FDs provide a way to represent the shape of a signal in the frequency domain[163]. Given a sequence of fixation points represented as coordinates over time, the Fourier Transform is applied to convert these coordinates from the time domain to the frequency domain. This transformation is beneficial because it allows us to capture periodic patterns and oscillations in the gaze data.

Given a sequence of fixation points (x_i, y_i) where $i = 0, 1, \ldots, N-1$, we treat x and y as discrete signals. The Discrete Fourier Transform (DFT) is applied to both x and y to obtain their frequency domain representations. For a discrete signal $x(n)$ with N samples, the DFT is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-i\frac{2\pi kn}{N}} \tag{4.4}$$

where $X(k)$ is a complex number comprising a real part $\text{Re}(X(k))$ and an imaginary part $\text{Im}(X(k))$:

$$\text{Re}(X(k)) = \sum_{n=0}^{N-1} x(n) \cdot \cos\left(\frac{2\pi kn}{N}\right) \tag{4.5}$$

$$\text{Im}(X(k)) = -\sum_{n=0}^{N-1} x(n) \cdot \sin\left(\frac{2\pi kn}{N}\right) \tag{4.6}$$

These components represent the cosine and sine components of the signal at each frequency k. The same approach applies to $Y(k)$.

Given the sampling frequency of 60 Hz and a time window of 5 seconds, the frequency range spans from -30 Hz to 30 Hz, with a frequency resolution of 0.2 Hz. We systematically vary the number of Fourier descriptors ($n_{\text{descriptors}}$) to identify the optimal balance between capturing key oscillatory patterns and maintaining model generalization. Lower values may retain only coarse movement trends, while higher values introduce finer details that may not improve classification. This exploration ensures a compact yet informative representation of fixation dynamics for robust recognition. Thus, the feature vector for each configuration is formed by concatenating the real and imaginary parts of the selected frequency components ($n_{\text{descriptors}} = k$) for both x and y coordinates:

$$\text{FD}_x = [\text{Re}(X(0)), \text{Im}(X(0)), \ldots, \text{Re}(X(k-1)), \text{Im}(X(k-1))] \tag{4.7}$$

$$\text{FD}_y = [\text{Re}(Y(0)), \text{Im}(Y(0)), \ldots, \text{Re}(Y(k-1)), \text{Im}(Y(k-1))] \tag{4.8}$$

The combined feature vector for both coordinates is:

$$\text{FD} = [\text{FD}_x, \text{FD}_y] \tag{4.9}$$

This feature vector provides a detailed frequency-based representation of the fixation points, highlighting essential periodicities and shape characteristics in the eye movement data.

Given the dataset **FD** of shape (m, n), where m is the number of samples and n is the number of features, PCA is applied to reduce dimensionality while preserving key variations[164]. It transforms the original descriptors into uncorrelated principal components—linear combinations of the original features—that capture the most significant data variations. The process involves:

1. Standardization: Subtract the mean of each feature and divide by its standard deviation to ensure all features contribute equally to the analysis.

2. Covariance Matrix Computation: Calculate the covariance matrix Σ to understand how the features vary together:

$$\Sigma = \frac{1}{m-1}\mathbf{FD}^T\mathbf{FD} \tag{4.10}$$

3. Eigenvalue Decomposition: Decompose the covariance matrix into eigenvalues and eigenvectors. The eigenvectors represent the directions of maximum variance, while the eigenvalues represent the magnitude of this variance.

4. Principal Components Selection: Initially, the top 3 eigenvectors corresponding to the largest eigenvalues are selected (i.e., $n_{\text{components}} = 3$) to form the principal components. This selection balances reducing dimensionality while retaining significant information. However, to fully explore the impact on classification performance, we also analyze the effects of selecting different numbers of principal components p.

The transformed dataset $\mathbf{PCA_{FD}}$ is given by:

$$\mathbf{PCA_{FD}} = \mathbf{FD} \cdot \mathbf{W} \tag{4.11}$$

where $\mathbf{W}$ is the matrix of the selected eigenvectors. By adjusting the number of retained components p, we create a compact representation of the data that highlights the most important variations in the gaze trajectories. Specifically, the transformed feature vector $\mathbf{PCA_{FD}}$ is:

$$\mathbf{PCA_{FD}} = [\mathbf{PCA_{FD1}}, \mathbf{PCA_{FD2}}, \ldots, \mathbf{PCA_{FDp}}] \tag{4.12}$$

These components represent the most significant variations in the gaze trajectories captured by the frequency domain features, capturing both the overall trajectory shape and the finer details of oscillatory behaviors.

4.3.4.3 Time-lagged correlation analysis

In the time domain, we analyze the horizontal (x) and vertical (y) coordinates of fixations into a complex signal (as shown in Equation 4.13) to effectively capture their synchrony, phase differences, and interactions. We extracted the features of time domain using autocorrelation [165]. Autocorrelation quantifies how a signal relates to a time-shifted version of itself, revealing repetitive patterns and periodicity in fixation trajectories. By computing autocorrelation in the complex domain, we can characterize both magnitude and directional consistency of movements over time, which would be difficult to achieve by treating x and y separately.

$$z(n) = x(n) + iy(n) \tag{4.13}$$

where i is the imaginary unit ($i^2 = -1$).

The autocorrelation of the complex signal $z(n)$ is defined as:

$$R_z(t) = \sum_{n=0}^{N-t-1} z(n) \cdot \overline{z(n+t)} \tag{4.14}$$

where $\overline{z}$ denotes the complex conjugate of z. This captures joint variations in x and y over time while preserving phase relationships.

To obtain a scale-invariant measure of periodicity, we normalize by the zero-lag value $R_z(0)$:

$$r_z(t) = \frac{R_z(t)}{R_z(0)} \tag{4.15}$$

The normalized autocorrelation coefficients $r_z(t)$ quantify repetitive movement patterns while ensuring comparability across different signal magnitudes.

To determine the optimal feature set, we evaluate $r_z(t)$ over a range of lag values (t) and select those maximizing classification accuracy across four detection types:

$$\text{Autocorr} = [r_z(1), r_z(2), \ldots, r_z(T)] \tag{4.16}$$

A complex signal inherently comprises a real and an imaginary part, each capturing different aspects of trajectory dynamics. To fully characterize these dynamics, we extract and analyze these components separately:

$$\begin{aligned}\text{Re}(r_z(t)) = &\frac{\sum_{n=0}^{N-t-1} \text{Re}(z(n)) \cdot \text{Re}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2} \\ &- \frac{\sum_{n=0}^{N-t-1} \text{Im}(z(n)) \cdot \text{Im}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2}\end{aligned} \tag{4.17}$$

$$\begin{aligned}\text{Im}(r_z(t)) = &\frac{\sum_{n=0}^{N-t-1} \text{Re}(z(n)) \cdot \text{Im}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2} \\ &+ \frac{\sum_{n=0}^{N-t-1} \text{Im}(z(n)) \cdot \text{Re}(z(n+t))}{\sum_{n=0}^{N-1} |z(n)|^2}\end{aligned} \tag{4.18}$$

Equations (4.17) and (4.18) decompose $r_z(t)$ into real and imaginary parts to capture different movement dynamics. The real part reflects synchrony between x and y movements at lag t, such as periodic scanning, while the imaginary part captures asymmetric interactions, like out-of-phase shifts, leveraging the phase information inherent in complex signals.

From the resulting real and imaginary parts of $r_z(t)$, we extract four features : their respective variances and maximum values.

$$\textbf{Autocorr}_{\textbf{Real}}\textbf{Max} = \max(\text{Re}(r_z(t))) \tag{4.19}$$

$$\textbf{Autocorr}_{\textbf{Imag}}\textbf{Max} = \max(\text{Im}(r_z(t))) \tag{4.20}$$

$$\mathbf{Autocorr_{Real}Var} = \mathrm{Var}(\mathrm{Re}(r_z(t))) \tag{4.21}$$

$$\mathbf{Autocorr_{Imag}Var} = \mathrm{Var}(\mathrm{Im}(r_z(t))) \tag{4.22}$$

The maximum values capture nonlinear peak correlations, while the variances summarize overall periodicity and interaction patterns, enhancing the representation of complex trajectory behaviors. These features provide insights into the visual tracking mechanisms underlying fixation dynamics.

4.3.5 Classification models

In this study, ten machine learning models were systematically evaluated for their effectiveness in detection type recognition. The basic classifiers included support vector machine (SVM) [166], known for handling high-dimensional spaces; k-Nearest Neighbors (kNN) [33], valued for its simplicity and pattern recognition; and multilayer perceptron (MLP), a neural network adept at capturing nonlinear relationships [123]. Advanced ensemble methods, including random forest [28], eXtreme Gradient Boosting (XGBoost), and gradient boosting decision tree (GBDT) and Light Gradient Boosting Machine (LightGBM), were also tested for their robustness and efficiency in complex datasets [124]. Additionally, three state-of-the-art deep learning models were incorporated: Transformer-MLP, which uses multi-head self-attention to encode fixation sequences and an MLP for classification, capturing long-range dependencies[81]; LSTM-Attn, a bidirectional LSTM with attention to model structured dependencies and highlight key fixation patterns[83]; and CNN-1D, employing 1D convolutional layers and max pooling to extract local gaze dynamics for hierarchical feature classification[84]. To ensure robust model evaluation and mitigate overfitting, we employed 5-fold cross-validation with an independent test set (20%), for generalization across all models. For deep learning models, we applied L2 regularization (weight decay = 1e-5) and Dropout (0.3) to reduce parameter overfitting, alongside an adaptive learning rate schedule (ReduceLROnPlateau) to optimize training. Additionally, the Synthetic Minority Over-sampling Technique (SMOTE) and Min-Max normalization were used to balance the dataset and standardize feature scales, minimizing bias and enhancing robustness [167]. The models were rigorously evaluated for classifying the four detection types, using accuracy for overall performance and recall for detecting critical failures, minimizing misses of subtle errors[120].

Furthermore, our proposed feature set was utilized to enhance the classification of four visual detection types, compared against twelve traditional eye-tracking features, matched to our proposed set's size for consistency. Traditional features include average fixation duration, standard deviation of fixation duration, fixation frequency, fixation count, saccade count, saccade frequency, blink rate mean, mean pupil size (left and right), fixation-saccade ratio, standard deviation of saccade magnitude, and maximum saccade magnitude, as extensively documented in previous research on eye movement analysis [16, 25, 75, 162].

4.4 Results

This section evaluates the effectiveness of our proposed hierarchical DF recognition framework through phase-specific state-of-the-art comparisons, as the framework addresses the spatiotemporal uncertainty of alerts and considers peripheral vision, making direct comparisons with pre-defined AOI-based frameworks challenging due to their inability to handle such randomness and visual factors. Detailed results follow.

4.4.1 Adaptive symbolic alert detection performance

In this study, we implemented the Adaptive Symbolic Alert Detection method to evaluate its effectiveness in accurately detecting and annotating alerts within an ATC simulation. To assess its performance, we collected 442 valid samples with effective eye-tracking and response data, as detailed in Fig. 6.1 in Appendix II. Our proposed method successfully detected and annotated alert regions in 421 of these samples, achieving an accuracy of 95.24%. This metric demonstrates the robustness of our method in accurately identifying alert regions within ATC radar-based alert scenarios, specifically ensuring the complete capture of aircraft icons and corresponding flight information. As a result, the 421 successfully annotated samples were selected for subsequent analysis.

Moreover, to compare our approach with state-of-the-art methods, we evaluated our Adaptive Symbolic Alert Detection method against the Segment Anything Model (SAM) [168] on the same dataset of 442 images. SAM is a general-purpose, transformer-based model designed for segmenting objects in diverse scenarios with high adaptability, which makes it a suitable benchmark for evaluating our adaptive approach. The comparison results show that our method showed better accuracy and efficiency. Specifically, our method correctly identified 421 segmented alert regions, compared to SAM's 405, resulting in a modest difference (95.24% vs. 91.63%). In terms of efficiency, our method achieved a total execution time of 2.883 seconds, averaging 0.007 seconds per image, while SAM required 14086.439 seconds total, averaging 31.870 seconds per image, highlighting a substantial processing time advantage for real-time applications. The SAM method's lower performance in domain-specific ATC applications stems from its general-purpose, transformer-based architecture. While it demonstrates high adaptability and robustness in diverse scenarios, its reliance on complex computations and a rigorous processing pipeline inherently compromises operational efficiency compared to our task-tailored methods [169]. For qualitative comparison, we present 12 sample images in Fig. 4.8, with 6 images each from our method and SAM, showcasing both consistent detections and notable differences.

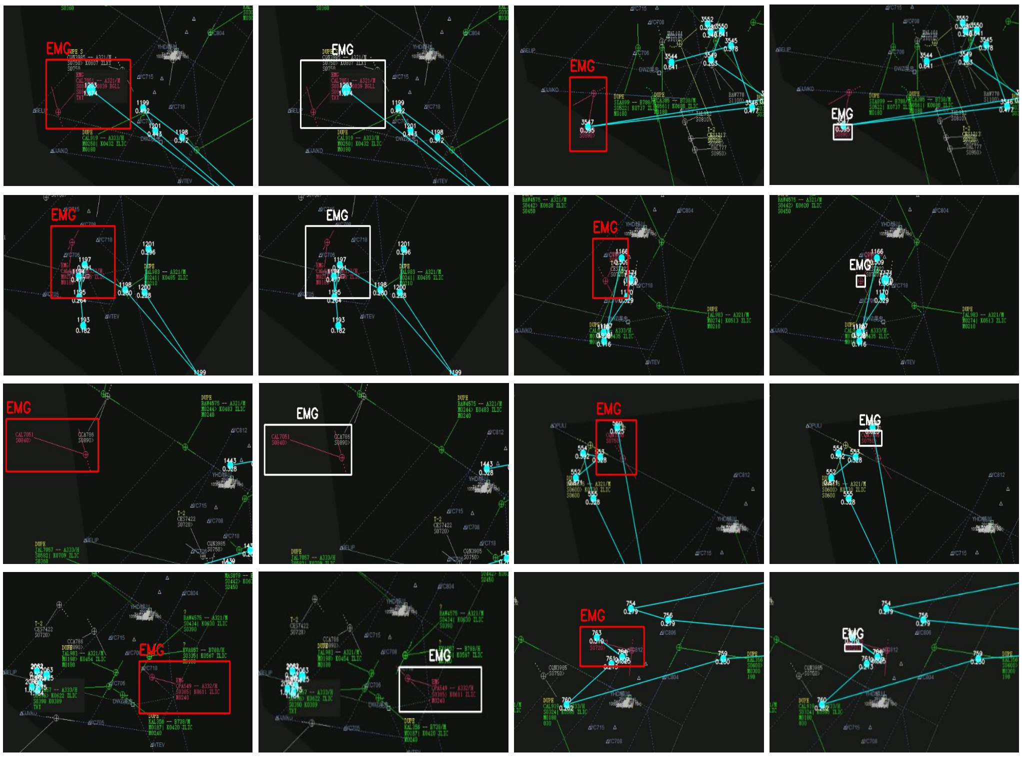


Figure 4.8: Qualitative comparison of the proposed Adaptive Symbolic Alert Detection method (red annotations) and SAM method (white annotations) on ATC alert-region detection, showing sample cases with both consistent and divergent outcomes.

4.4.2 Visual attention engagement assessment

An LLM-based visual processing approach was used to classify visual attention engagement on annotated alerts into Captured and Non-captured alerts. A sample output illustrating this process is provided, as shown in Appendix 6.2. The 421 successfully annotated alert samples from the previous section were categorized into four expert-derived ground truth labels: Effective Detection (291 samples), Peripheral Vision Neglect (52 samples), LBFTS Error (21 samples), and Pre-attentive Detection (57 samples). These labels serve as the reference standard for evaluating the LLM's accuracy in processing visual attention patterns. As shown in Fig. 4.9, the LLM's evaluations were divided into Captured and Non-captured alerts, further distinguishing correct and incorrect assessments. Among the 312 Captured alert cases, GPT-4o correctly identified 231 instances (74.04%), while 81 cases were misclassified. For the 109 Non-captured alert instances, GPT-4V correctly assessed 61 cases (55.96%) but misclassified 48 cases.

The results indicate that GPT-4V performs more effectively in scenarios where alerts were captured, likely due to clearer fixation overlaps with alert regions, reflecting its strength in processing straightforward visual inputs. However, its performance declines in Non-captured scenarios, which are inherently more challenging due to the complexity and ambiguity of peripheral or erratic eye movement trajectories, sample imbalance (312 Captured vs. 109 Non-captured alerts), and limitations in LLM' pretraining on general-purpose images, underrepresenting dynamic ATC contexts. This disparity suggests that while GPT-4V demonstrates potential for processing captured alerts, its accuracy in ambiguous, Non-captured cases, such as Peripheral Vision Neglect, requires refinement. Overall, these findings underscore the potential of integrating LLM-based approaches with our advanced multi-domain detection algorithms, enhancing alert detection systems in complex operational environments like ATC.

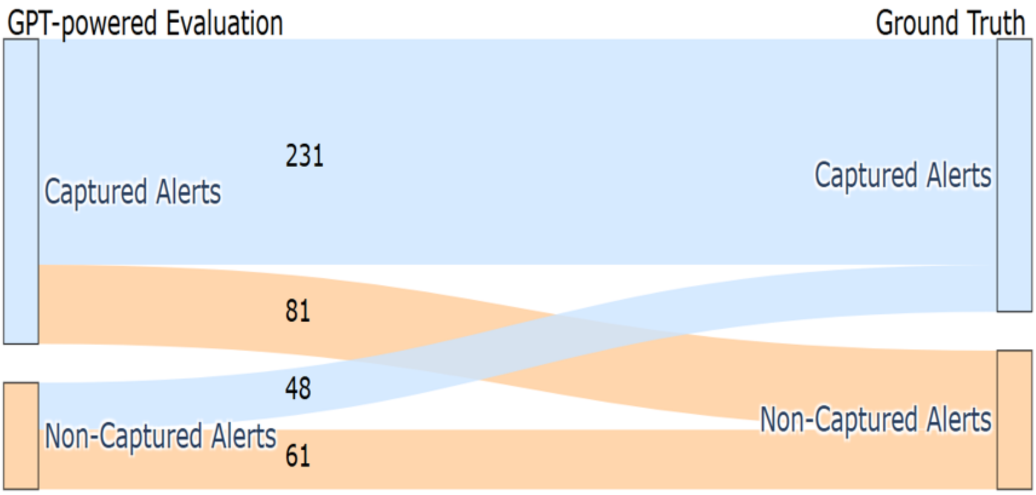


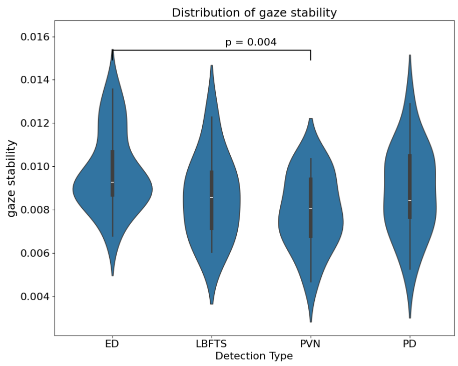
Figure 4.9: Performance of LLM (GPT-4V) evaluation of visual attention to alerts, compared against expert-derived ground truth, showing correct and incorrect classifications.

4.4.3 Domain-specific performance in distinguishing DF

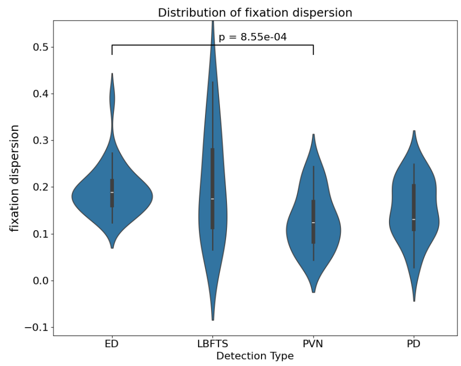
4.4.3.1 Spatial analysis

The analysis of multi-dimensional spatial features across four detection types provides crucial insights into how different visual detection types are characterized by specific eye movement patterns. To ensure a more reliable and representative comparison, the Mann-Whitney U test was conducted based on the average values of each spatial feature for each participant, rather than using the overall sample size. This participant-level analysis reduces the influence of intra-individual variability, offering a more accurate assessment of between-group differences. The

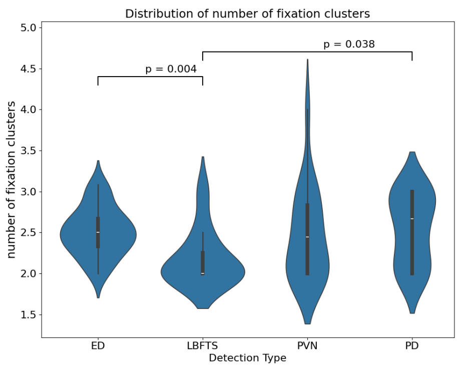
Mann-Whitney U test, a non-parametric method, was chosen for its robustness in handling non-normal data and effectiveness in comparing independent groups [170]. The four spatial features examined reveal distinct patterns among the detection types, particularly highlighting statistically significant differences, as shown in Fig. 4.10.



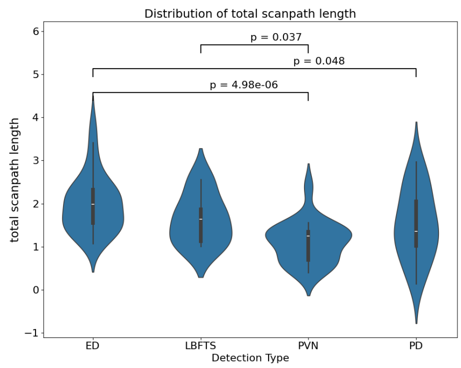
(a) Gaze stability



(b) Fixation dispersion



(c) Number of fixation clusters



(d) Total scanpath length

Figure 4.10: Spatial feature analysis across four detection types, with significant differences via Mann-Whitney U test. (Note: Horizontal lines indicate significant differences between pairs, with p-values shown above each line.)

In analyzing visual scanning behavior across different detection types, several key patterns emerge across the subfigures. In subfigures (a) and (b), the analysis of gaze stability and fixation dispersion reveals consistent patterns between ”Peripheral Vision Neglect" and "Effective Detection." "Peripheral Vision Neglect" exhibits a significantly more concentrated distribution of gaze points within fixations (p = 3.939e-03) and lower fixation dispersion (p = 8.547e-04) compared to "Effective Detection." These findings indicate a more concentrated and stable visual focus during "Peripheral Vision Neglect," suggesting that participants' intense focus on central stimuli may limit their ability to effectively track and respond to peripheral cues, leading to the neglect of peripheral stimuli. In subfigure (c), the number of fixation clusters shows significant differences between "LBFTS" errors and both "Effective Detection" p = 3.800e-03) and "Pre-attentive Detection" (p= 3.843e-02), with "LBFTS" errors having fewer clusters. This suggests that during LBFTS errors, although the alert is perceived, the limited fixation clusters indicate insufficient cognitive resource distribution for effective processing and interpretation. This cognitive disconnect leads to an inadequate response despite initial detection. In subfigure (d), the analysis of total scanpath length reveals significant differences between "Peripheral Vision Neglect" and both "LBFTS" errors (p = 3.683e-02) and "Effective Detection" (p = 4.981e-06). "Peripheral Vision Neglect" shows the shortest scanpath length overall, indicating that participants' eye movements are most constrained during this condition. This suggests a highly focused visual attention on specific tasks, leading to the neglect of peripheral alerts. In addition, the significant difference between "Pre-attentive Detection" and "Effective Detection" (p = 4.753e-02), suggests that when participants perceive an alert peripherally and respond without direct fixation, their visual search is more efficient, requiring less extensive eye movement. This efficiency likely stems from quickly detecting and processing the alert within peripheral vision, reducing the need for broader scanning typical in "Effective Detection".

Collectively, these findings across the subfigures underscore the distinct visual scanning behaviors and the importance of spatial features associated with different detection types, providing valuable insights into the underlying mechanisms that contribute to detection successes and failures.

4.4.3.2 Frequency domain analysis

This section evaluates the classification performance of frequency-domain features in distinguishing peripheral vision-aware DF in ATC. By isolating Fourier Descriptors (FDs) and PCA components, we assess their contributions to identifying four visual detection types. Random Forest, the best-performing model among ten classifiers, achieved the highest test precision across all configurations.

Fig. 4.11 illustrates the classification accuracies for FD (k) ranging from 3 to 10 and PCA components (p) from 2 to 4, selected to balance capture sufficient frequency detail while avoiding noise in Discrete Fourier Transforms, with PCA limited to prevent redundancy. In general, accuracy peaks at 82% with k = 5 and p = 4, suggesting that this combination optimally balances the richness of Fourier descriptors and the reduction in dimensionality provided by PCA. Interestingly, the middle values of k (e.g. 6) with lower p (e.g., 2) tend to result in lower precision (61%), which could be due to the descriptors in these values not providing enough variation or being less representative of the shape characteristics needed for optimal classification. In contrast, lower values of k (e.g., 3) also lead to reduced accuracy (64%), likely due to insufficient descriptor complexity. The trend suggests that increasing k improves accuracy up to a point, but the model benefits the most with the p = 4 components for the reduction of dimensionality, as higher k without appropriate adjustments p does not consistently improve performance. These findings highlight the importance of carefully tuning FD and PCA parameters to optimize frequency feature extraction for complex classification tasks, revealing their role in capturing nuanced gaze dynamics critical for ATC detection. To further explore the frequency domain's role across detection types,

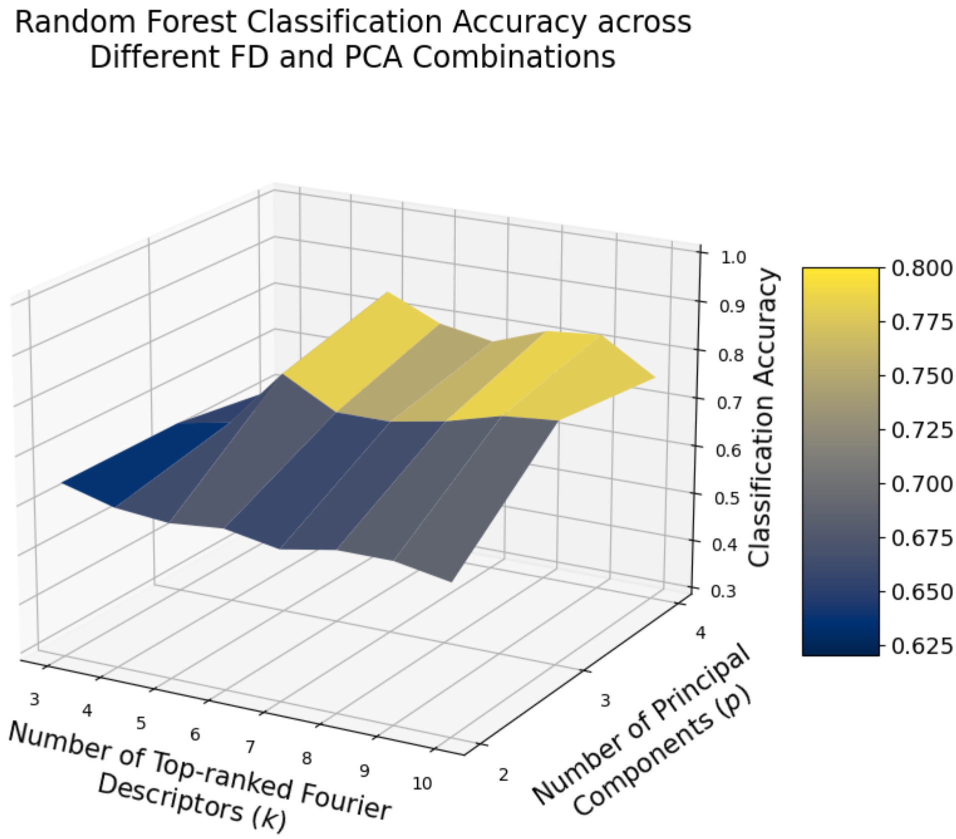


Figure 4.11: Random forest classification accuracy for varying Fourier Descriptors (k=3–10) and PCA components (p=2–4), peaking at 81% with k=5 and p=4 for four visual detection types.

we examined the per-class performance of the optimal configuration (k=5, p=4) using RF, as shown in Fig. 4.12. The normalized confusion matrix illustrates the classification performance, with the diagonal values representing the proportion of correctly classified samples for each detection type: Effective Detection (0.72),

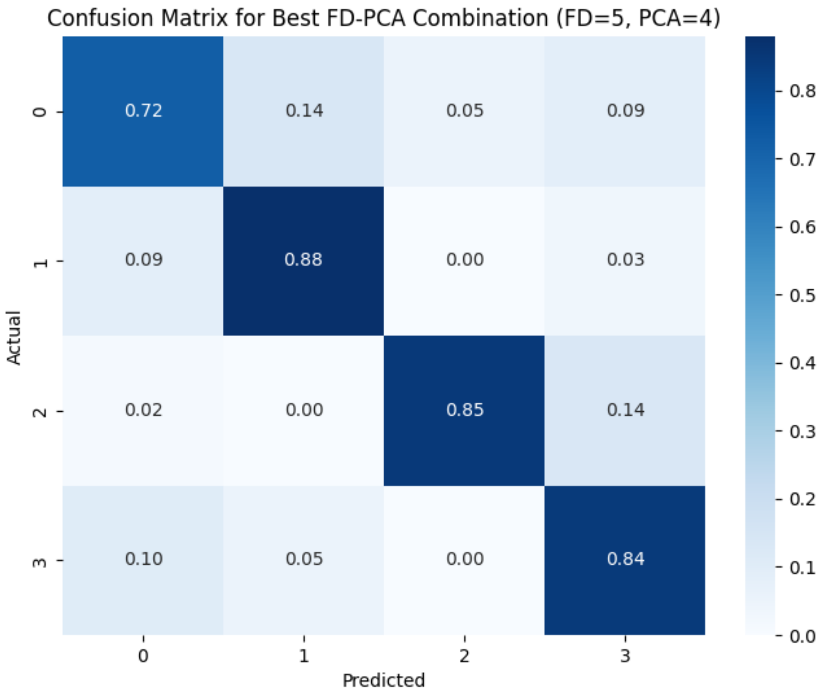


Figure 4.12: Confusion matrix for the random forest model using four frequency-domain features, showing performance across four detection types (0: Effective Detection, 1: Peripheral Vision Neglect, 2: LBFTS Error, 3: Pre-attentive Detection)

Peripheral Vision Neglect (0.88), LBFTS Error (0.85), and Pre-attentive Detection (0.84). Additionally, the recall scores were calculated for each class: Effective Detection (0.60), Peripheral Vision Neglect (0.88), LBFTS Error (0.93), and Pre-attentive Detection (0.81). These results demonstrate that frequency features excel at detecting Peripheral Vision Neglect and LBFTS Errors, likely due to their sensitivity to rapid, oscillatory saccades and constrained fixations, respectively, as supported by prior research on saccadic dynamics [161]. However, the lower performance for Effective Detection suggests integrating temporal or spatial features to enhance its detection. Overall, the findings highlight the frequency domain's ability to model rhythmic gaze disruptions, enhancing ATC safety by identifying subtle cognitive lapses and attentional constraints with distinct dynamic patterns.

4.4.3.3 Time-lagged correlation analysis

This section examines the classification performance of time-lagged autocorrelation features, evaluating their role in distinguishing four DF types in ATC. By isolating features capturing maximum autocorrelation values and their variances of the real and imaginary parts across lags 2–10, we clarify their contribution to recognizing proposed four visual detection type. Random Forest, the highest-performing model among ten classifiers, achieved the best test accuracy across all configurations. Fig. 4.13 presents a radar plot of test accuracies for lags ranging from 2 to 10, selected based on typical fixation counts (13–22 per sample) to ensure meaningful temporal patterns without overcomplicating the model. Results show accuracy peaks at 81% for lag=9, with a range from 72% (lag=2) to 80% (lag=8). This trend indicates that longer lags, particularly lag=9, capture sustained attention patterns critical for ATC detection, revealing their role in modeling temporally distinct gaze behaviors.

To evaluate the performance of time domain features across detection types, we analyzed the per-class performance of the optimal configuration (lag=9) using Random Forest, as shown in Fig. 4.14. The normalized confusion matrix reveals classification performance, with diagonal values indicating correct classification rates: Effective Detection (0.59), Peripheral Vision Neglect (0.86), LBFTS Error (0.92), and Pre-attentive Detection (0.84). The model excels in classifying LBFTS Error and Peripheral Vision Neglect, reflecting time-lagged features' ability to capture distinct temporal dynamics, such as centrally focused fixations in

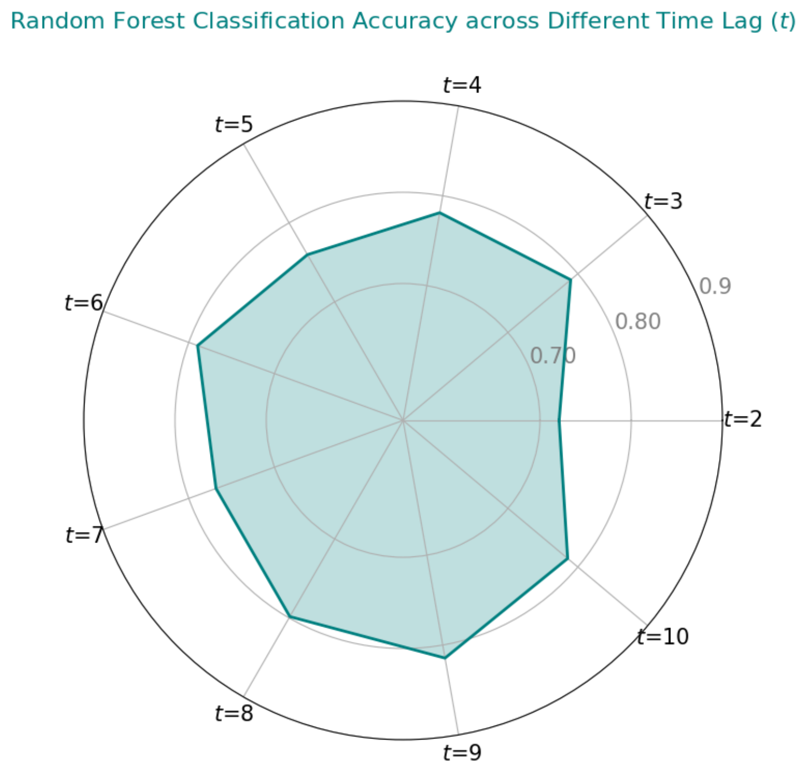


Figure 4.13: Random forest classification accuracy across time lags (t=2–10) in autocorrelation analysis, with peak performance of 81% at t=9 for four visual detection types.

LBFTS and regular, concentrated patterns in PVN, critical for identifying these failures. However, the lower performance for Effective Detection suggests integrating frequency or spatial features to enhance its detection.

To quantify these temporal dynamics, we analyzed the average maximum autocorrelation values at lag=9. Peripheral Vision Neglect and Pre-attentive Detection exhibit the highest real part maximum values (0.86 each), indicating strong synchronization in fixation coordinates, with low imaginary part values (0.03 and 0.02) suggesting minimal phase differences. LBFTS Error (real 0.83, imag 0.03) and Effective Detection (real 0.82, imag 0.05) show slightly lower synchroniza-

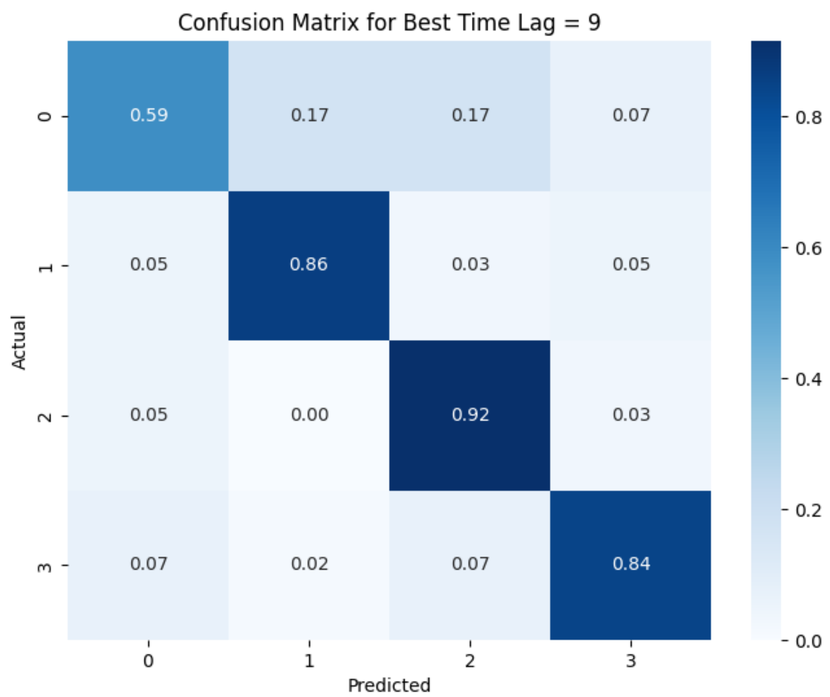


Figure 4.14: Confusion matrix for the random forest model using four time-domain features, showing performance across four detection types (0: Effective Detection, 1: Peripheral Vision Neglect, 2: LBFTS Error, 3: Pre-attentive Detection)

tion. Despite Peripheral Vision Neglect's strong synchronization, LBFTS Error's superior performance suggests time-lagged features better capture its centrally focused fixation patterns, possibly due to distinct temporal consistency in gaze concentration. For qualitative comparison, Fig. 4.15 displays sample trajectories (top-ranked by autocorrelation values at time-lag=9), revealing prolonged, inconsistent temporal shifts in Effective Detection, rhythmic and sustained fixation cycles in PVN, stable and prolonged central focus in LBFTS, and variable temporal patterns with intermittent attention shifts in Pre-attentive Detection, validating the distinct time-lagged behaviors across categories.

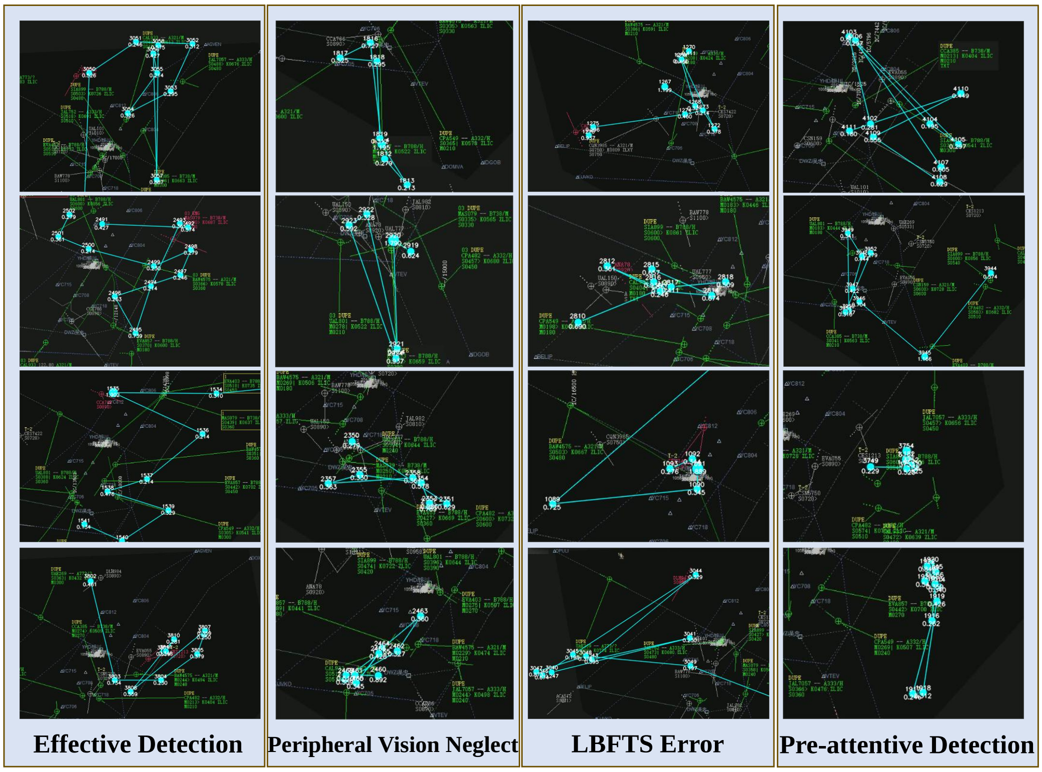


Figure 4.15: Qualitative comparison of temporal characteristics across four detection types using samples with top-ranked maximum autocorrelation samples.

4.4.4 Integrated domain performance in visual detection failure recognition

This study evaluates the classification performance of visual DF recognition, focusing on four distinct detection types, by comparing the proposed fixation coordinate-sensitive multi-domain feature set with several alternatives. The proposed feature set integrates 12 features across three domains—four spatial, four frequency (FD=5, PCA=3, the optimal configuration identified in Section 4.4.3.2), and four temporal (lag=9, the optimal value identified in Section 4.4.3.3)—to capture a comprehensive representation of gaze behavior. For comparison, we include a traditional feature set (also containing 12 features to ensure fairness), alongside domain-specific subsets: single-domain sets (spatial-only, frequency-only, temporal-only, each with four features) and dual-domain combinations (spatial-frequency, spatial-temporal, frequency-temporal).

Table 4.1: Accuracy (%) of traditional and the proposed fixation coordinate-sensitive multi-domain feature sets for four-category detection-type recognition across ten classifiers via ablation analysis of the proposed set with single- and dual-domain tests.

Feature Set	SVM	kNN	MLP	RF	XGBoost	GBDT	LightGBM	Transformer-MLP	LSTM-Attn	CNN-1D
Traditional features set	66.09	79.40	78.11	82.40	82.40	76.82	**83.69**	82.83	79.40	64.38
The proposed feature set	79.83	81.55	85.84	90.13	88.41	82.83	92.27	**93.13**	88.84	74.68
Spatial-frequency features	71.24	80.26	77.68	85.84	84.55	76.39	87.12	**90.30**	69.50	86.27
Spatial-temporal features	64.38	76.82	75.11	**83.26**	81.12	75.11	**83.26**	82.00	67.81	79.40
Frequency-temporal features	69.53	80.26	79.40	87.55	87.55	81.97	**90.99**	89.00	68.00	80.26
Spatial features-only	54.08	72.10	54.94	**76.82**	69.96	63.52	71.24	71.24	54.94	51.07
Frequency features-only	62.23	71.67	60.09	**80.69**	77.25	70.82	80.26	76.82	61.37	66.09
Temporal features-only	57.51	69.10	57.51	**81.00**	79.40	71.24	78.97	71.67	57.51	60.94

Table 4.1 presents the classification accuracy across ten classifiers, highlighting the robustness of the proposed multi-domain feature set through an ablation study. Firstly, the proposed set consistently outperforms the traditional feature set across all models, achieving the highest accuracy of 93.13% with Transformer-MLP, compared to 82.83% for the traditional set—a 10.30% improvement. LightGBM ranks second at 92.27% (vs. 83.69% for traditional features), demonstrating the proposed set's superior generalization across classifiers. Ablation analysis further reveals the contribution of individual and paired domains. Single-domain subsets perform significantly worse, with spatial features-only yielding the lowest accuracy (peak at 76.82%), followed by temporal features-only (peak at 81.00%) and frequency features-only (peak at 80.69%), underscoring the limitations of relying on isolated domains. Dual-domain combinations, however, show marked improvements over single domains: frequency-temporal features achieve 90.99% with LightGBM, and spatial-frequency features reach 90.30% with Transformer-

MLP. Notably, spatial-temporal features underperform (peak at 83.26%), suggesting potential challenges in model convergence for this combination. These results confirm the necessity of integrating all three domains, as the proposed multi-domain set consistently surpasses both single- and dual-domain subsets, highlighting the critical role of domain interplay in enhancing detection-type recognition accuracy.

4.5 Discussion

This study establishes a novel peripheral vision-aware visual detection classification and validates it by introducing a Hierarchical Detection Failure Recognition Framework, demonstrating the distinct differences between various detection types. We first conducted an ATC simulation experiment to provide empirical data for the framework analysis. The first phase of the framework introduces the Adaptive Symbolic Alert Detection method, which identifies alert regions with spatiotemporal uncertainty on the ATC radar interface. This method enables robust identification of alert information on radar interfaces, achieving a 95.24% precision rate. Our study, aligned with prior research isolating critical events to establish a foundational understanding of psychophysiological patterns [58, 153], currently focuses on single-alert scenarios. Should overlapping alert scenarios be further explored, our method effectively identifies the overall alert regions; however, future research should focus on delineating individual alert boundaries within overlaps and incorporating semantic information and alert prioritization to enhance its applicability in complex ATC environments.

In the second phase, GPT-4V is employed to evaluate visual attention to alerts, facilitating the classification of samples into four proposed detection types. Our findings reveal that the LLM's evaluation of detected alerts shows a clear advantage in scenarios where alerts were captured accurately, with a correct evaluation rate of 74.04%. This outcome aligns with findings in similar studies, such as the research of [80], where LLMs were integrated with EEG and eye-gaze data, achieving a classification accuracy of over 60%. These studies collectively highlight the potential of LLMs in tasks requiring semantic and contextual understanding, especially when the visual or linguistic data is straightforward and less ambiguous. However, the LLM's performance demonstrates a need for further refinement in more challenging situations, such as the evaluation of Non-captured alerts where the correct rate dropped to 55.96%.

A potential explanation of the disparity in performance between Captured and Non-captured alert evaluations can be attributed to three key factors. The complexity and ambiguity of Non-captured scenarios, where eye movement trajectories often exhibit peripheral or erratic patterns, challenge GPT-4V's visual capabilities, as these patterns reflect peripheral vision or cognitive failures (e.g., attention dispersion, lack of understanding). Second, sample imbalance—312 Captured versus 109 Non-captured alerts—biases LLM toward captured scenarios, consistent with AI bias studies[171], which highlight reduced sensitivity to rare events. Third, GPT-4V's pretraining on general-purpose images and texts underrepresents dynamic ATC contexts, limiting its ability to capture subtle differences in peripheral vision or unstable trajectories[172]. To address the challenges, we propose building on the results of this study as a baseline to explore different LLM models with prompt designs, such as chain-of-thought (CoT) prompting for reasoning about gaze patterns or instruction-based prompting with task-specific guidelines for peripheral vision dynamics, and fine-tuning LLMs with ATC-specific datasets. This approach aligns with recent advances in LLM optimization[80] and aims to enhance the model's ability to capture nuanced eye movement patterns, ultimately improving performance in ATC alert detection tasks.

Additionally, the framework introduces a novel fixation coordinate-sensitive multi-domain feature set, achieving 93.13% accuracy in peripheral vision-aware detection, surpassing traditional eye-tracking metrics (83.69%) and both single- and dual-domain features. Compared to prior research on four visual attentional type classification[161], which achieved only 79% accuracy, our results highlight the significant advantage of fixation coordinate-sensitive multi-domain feature integration in capturing nuanced visual tracking dynamics. The following discussion will examine how the four detection types are differentiated through the employed domain-specific analysis, shedding light on the underlying mechanisms and their potential implications for improving detection accuracy.

In this study, spatial feature analysis offers critical insights into how visual attention is distributed across detection successes and failures. Previous studies, such as [147], have suggested that efficient eye movement patterns are typically characterized by broader and more extensive tracking trajectories, indicating effective visual processing. Consistent with these findings, our study found that LBFTS errors are associated with significantly fewer fixation clusters compared to ”Effective Detection” and ”Pre-attentive Detection”. Additionally, the significantly higher gaze stability and lower fixation dispersion observed in ”Peripheral Vision Neglect” compared to ”Effective Detection” align with the challenges in maintaining consistent attention to peripheral stimuli, as highlighted by [173]. Thus, the two DF result from difficulties in processing peripheral information, reflecting a more focused but less exploratory search strategy. Furthermore, the two DF types are significantly different, as reflected by ”Peripheral Vision Neglect” showing significantly shorter scanpath lengths than ”LBFTS,” indicating over-focus on specific tasks and neglect of peripheral alerts. Besides, the two successful detection types also differ significantly, with ”Pre-attentive Detection” exhibiting shorter lengths than ”Effective Detection,” indicating a more efficient, less extensive search during peripheral detection without direct fixation.

The introduction of frequency domain analysis has proven to be a powerful tool for distinguishing between different detection types by analyzing how varying the number of FDs and PCA components impacts recognition accuracy. Previous studies, such as the examination of saccadic eye movements in [163], have highlighted the importance of FDs in capturing dynamic visual patterns. Similarly, [174] demonstrated that incorporating frequency domain features can significantly enhance temporal information processing in emotion recognition tasks. In our study, we found that increasing the FD generally improves accuracy, with a peak observed at FD=5 and PCA=4, achieving an accuracy of 81%. Furthermore, the strong performance in detecting LBFTS Errors and Peripheral Vision Neglect underscores the frequency domain’s sensitivity to rhythmic gaze disruptions, such as oscillatory saccades and constrained fixations, offering a reliable approach to identify critical cognitive lapses in high-stakes ATC settings. However, its limitation in capturing dispersed gaze patterns, as observed in Effective Detection, underscores the need for multi-domain frameworks integrating time, frequency, and spatial features. Additionally, while our noise mitigation strategies minimize distortion, advanced denoising techniques like wavelet transforms or adaptive filtering could further suppress high-frequency jitters and stabilize visual trajectory patterns, addressing residual noise concerns[175].

Building on the insights from the time-lagged correlation analysis, this discussion deepens the understanding of temporal dynamics across detection types, with the highest accuracy 81% at lag=9 underscoring the importance of capturing sustained attention patterns essential for identifying cognitive lapses such as LBFTS Errors and Peripheral Vision Neglect. This aligns with [135], which emphasizes the critical role of temporal dependencies in visual attention mechanisms, a finding reinforced by [176], who linked LBFTS failures to ineffective temporal processing. The robust performance in detecting centrally focused fixations and rhythmic gaze disruptions further validates a strong framework for pinpointing high-risk attentional failures; however, the difficulty in capturing dispersed patterns in Effective Detection highlights the necessity for adaptive multi-domain approaches. These observations inspire future research into dynamic lag optimization, real-time integration of temporal features with spatial and frequency domains, and personalized gaze modeling to accommodate diverse attentional dynamics, thereby enhancing the resilience of automated ATC systems by adapting to individual operator variability and minimizing detection errors.

The proposed hierarchical peripheral vision-aware framework enhances ATC as a critical transportation system by dynamically assessing operator attentional states to improve alert systems, effectively differentiating meaningful alert engagement from mere gaze direction—a distinction vital in ATC where missed warnings stem not from lacking visual contact but from attentional prioritization failures[161]. Building on this, the framework advances beyond immediate detection by fostering an intelligent, human-centered ATC support system that identifies DF and informs adaptive intervention strategies. By analyzing visual attention fluctuations in dynamic air traffic scenarios, it enables personalized alert optimization, adjusting warning intensity and format to individual controller engagement patterns. This shift from reactive monitoring to proactive enhancement strengthens ATC resilience against attentional lapses, preserving situational awareness and decision-making integrity in high-stakes settings. Future research should focus on practical deployment strategies and real-time implementation feasibility within ATC centers.

4.6 Concluding remarks

This study introduced a novel classification framework for visual detection results incorporating peripheral vision awareness and human attentional states, highlighting a broader and more structured range of visual DF. Furthermore, a hierarchical framework has been developed to effectively recognize the proposed DF types in ATC-specific environments. First, the proposed Adaptive Symbolic Alert Detection method successfully addressed the spatiotemporal uncertainties of alert appearances on ATC radar interfaces, achieving a precision rate of 95.24%. Additionally, the GPT-4V-powered LLMs showed significant potential in enhancing the verification of controllers' visual attentional engagement with alerts, extending the capabilities beyond what traditional eye-tracking methods alone could achieve. Finally, the proposed fixation coordinate-sensitive multi-domain feature set markedly enhanced detection-type recognition, with accuracy rising from 83.69% (traditional features) to 93.13% with Transformer-MLP, outperforming single- and dual-domain representations and other models, demonstrating the synergistic effectiveness of multi-domain feature integration.

Chapter 5

Fixation state-enabled LLM-enhanced DF prediction

In this chapter, a novel approach is proposed for predicting DF by leveraging fixation-state transition features and L-PTT method. The study focuses on capturing pre-alert attentional patterns and enhancing prediction interpretability through 2D heatmap analysis. The structure of this chapter is organized as follows: Section 5.1 presents the research motivation and challenges in DF prediction. Section 5.2 details the methodology, including pre-alert period identification, fixation-state transition feature extraction, and the L-PTT framework. Section 5.3 evaluates the proposed feature set, analyzes L-PTT optimization, and compares performance with state-of-the-art methods. Sections 5.4 and 5.5 discuss key findings, practical implications, and offer concluding remarks.

5.1 Background

5.1.1 Motivation

As ATC demands precise and proactive responses to dynamic alerts, predicting and mitigating these failures is essential to enhancing system reliability and preventing accidents. Research on attentional lapses has significantly advanced the understanding of DF prediction [21, 40, 45]. Foundational studies by Simons and Chabris [9, 10], and Mack and Rock [11, 12] illustrate mechanisms like IB and LBFTS errors, highlighting the role of cognitive overload and expectancy in missing critical alerts. Building on this understanding, DF has been systematically classified into categories such as ordinary blindness, LBFTS errors, and misinterpretation, applying situational awareness theory to frame how cognitive processes contribute to such failures [161]. Based on these insights, empirical research [18, 25, 114] further highlights the role of eye-tracking techniques in fatigue and attention lapses prediction, employing metrics like fixation patterns, pupil dynamics, and scanpaths, offering non-invasive tools for detection failure analysis. Moreover, studies leverage practical algorithms, such as LSTM, random forest, and transformers, to model failure patterns across safety-critical domains, enhancing prediction capabilities [79, 90]. Additionally, our prior work advances this by incorporating peripheral vision to investigate the scope of visual attention, distinguishing DF from complex see-without-looking situations, proposing LLM-based fixation-coordinate-sensitive analysis for robust DF recognition [177]. These studies collectively offer robust strategies for predicting DF.

5.1.2 Challenges

While prior studies have developed predictive methods for some cognitive activities (e.g., workload, stress, and fatigue), specialized predictive approaches for DF remain challenging due to their unique characteristics [25, 28]. On the one hand, DF prediction involves dynamic attention-switching patterns prior to alert triggering, differing from post-hoc identification that analyzes gaze trajectories after stimulus appearance [18, 178]. This poses a primary challenge in the complexity of feature representation for accurate prediction. For instance, Verschooren et al. [179] emphasize that attention-switching between internal and external information involves asymmetrical switch costs, suggesting that dynamic transitions in attention allocation are resource-intensive and context-dependent. These findings highlight the need for features that model the dynamic interplay of attention transitions to enable early detection of potential failures [20].

On the other hand, the low-frequency nature of DF leads to prediction classifiers inherently biased toward majority classes, resulting in poor recall rates for true DF instances [21, 39]. To mitigate this imbalance, recent studies have employed Bayesian optimization methods for algorithmic hyperparameter tuning [91, 180]. Beyond algorithmic optimization, standard practices also focus on data representation and decision boundaries to improve model performance. For instance, data resampling techniques like SMOTE augment minority class representation by generating synthetic samples, rebalancing the dataset [87, 181]. These efforts are further reinforced by post-processing methods, such as probability calibration and probability threshold adjustment, to refine classifier outputs [93]. Established tools like grid search [182], PR curves [183], and ROC analysis [94]optimize thresholds for rare events to maximize their recall rates and F1-scores. However, a key limitation persists: static thresholds cannot adapt to sample-specific feature variations [184]. Key feature value ranges strongly correlate with minority DF type prediction probabilities [129, 185], and identifying these relationships enables adaptive probability thresholds to lower decision boundaries for relevant DF samples, boosting DF type recall. This need for intelligent threshold adjustment has driven interest in generative AI. Notably, LLMs have shown promising solutions in analyzing complex patterns in visual and textual tasks such as accident analysis and reasoning, human visual attention analysis [98, 186], as shown in our prior work on assessing visual engagement on alerts [177]. By learning intricate feature-conditioned relationships in imbalanced data, LLMs enable intelligent threshold adaptation, enhancing the reliability of DF prediction compared to conventional methods.

5.1.3 Objective

Therefore, this study aims to address the above challenges for robust DF prediction. Firstly, a fixation-state transition feature set is proposed to capture human attentional switching patterns, which focuses on capturing switches among distinct fixation states, enhanced by state-specific frequency metrics and spatial sequence patterns [71]. Moreover, we develop L-PTT method to enable intelligent threshold adaptation in DF prediction. The LPT method operates through an LLM that: (1) establishes initial probability thresholds by analyzing predictive probability distributions in non-DF and DF classes, and (2) generates adaptive threshold adjustment rules by learning feature-probability relations. This allows the prediction model to automatically calibrate class-specific probabilistic thresholds according to evolving feature characteristics, significantly enhancing detection sensitivity for critical DF events while preserving overall stability. Moreover, we extend the L-PTT method to investigate: (i) how input feature formats (2D heatmaps vs. numerical table) affect LLM-driven threshold accuracy, and (ii) the effectiveness of applying LLM-generated threshold plans via Exhaustive tuning (all adjustments) versus Priority-based tuning (high-impact adjustments first). These investigations advance our understanding of optimal LLM integration for adaptive prediction systems.

5.2 Methodology

This section presents the methodology for specialized DF prediction, integrating three coordinated components, as shown in Fig. 5.1. First, construct a feature set from eye movement data to characterize pre-alert attentional patterns, quantifying state transition dynamics and their spatial correlates. Second, perform DF prediction using the LLM-driven probabilistic threshold tuning (L-PTT) method, leveraging LLMs to enhance prediction reliability through three key operations: learning feature-probability relationships from train data distributions to initialize probabilistic thresholds and generate threshold adjustment rules, comparing 2D heatmap and numerical input formats to determine optimal feature representations, and implementing dynamic threshold adjustments through either exhaustive or priority-based tuning strategies. Finally, a comprehensive evaluation assesses system performance through class-specific metrics. Together, these components form an adaptive framework that continuously refines DF predictions based on evolving attentional patterns.

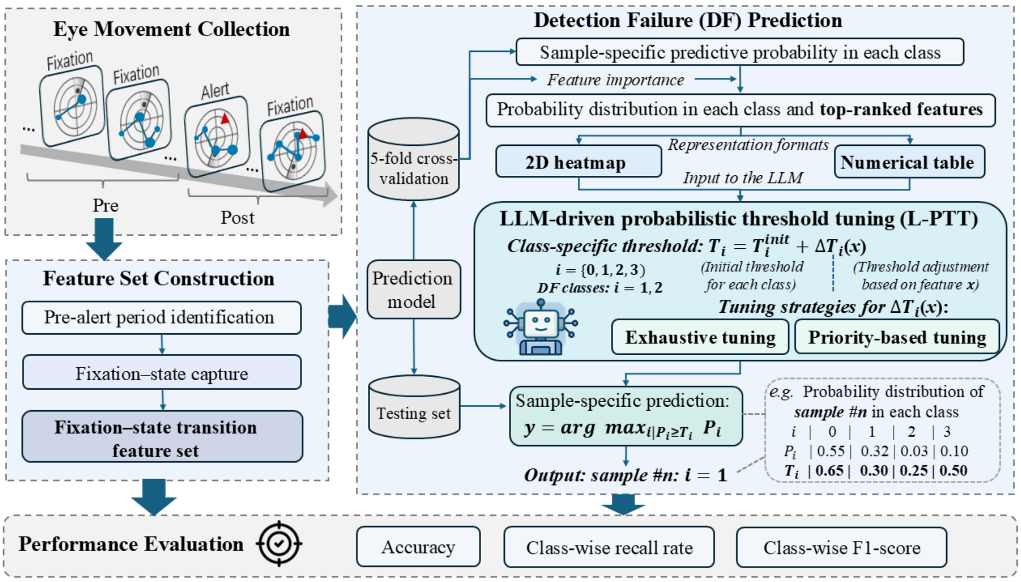


Figure 5.1: Methodology overview: an integrated pipeline for DF prediction through fixation-state transition feature set and LLM-driven probabilistic threshold tuning.

5.2.1 Feature set construction

Under feature set construction, pre-alert period identification defines the pre-trigger temporal window, fixation-state capture records eye movement states, and fixation-state transition feature set quantifies dynamic shifts and spatial correlates. Below is a detailed explanation.

5.2.1.1 Pre-alert period identification

The eye-tracking data originates from a controlled ATC simulation experiment in our prior work [177], capturing 421 warning-response instances classified into four behavioral categories: two non-DF types—NDF-1 (effective detection, 291 cases) and NDF-2 (pre-attentive detection, 57 cases)—and two DF types: DF-1 (peripheral vision neglect, 52 cases) and DF-2 (LBFTS errors, 21 cases). Focusing on the four detection types, we analyze attentional patterns within pre-defined temporal windows preceding warnings. The analysis adopts a 5-second baseline window (0 to -5s relative to warning onset), a duration empirically validated across ATC simulations and eye-tracking research as effectively capturing the eye movement trajectory patterns of different DFs [177]. To address severe class imbalance (e.g., only 21 DF-2 cases), we employ a sliding window approach with 1-second steps to augment samples while maintaining temporal relevance [75, 149]. The 1-second step size is physiologically justified by the 60Hz eye-tracking sampling rate (16.7ms/frame), ensuring each window captures new gaze data without requiring lengthy pre-alert observation periods. Based on the above settings, we evaluate three distinct pre-alert period configurations: Focused period: 0 to -5s, 1×5s window; Moderate period: 0 to -9s, 5×5s windows; Extended period: 0 to -14s, 10×5s windows. This stratified design provides progressive sample expansion (1× → 5× → 10×) while preserving the 5s attentional analysis window established in ATC studies.

5.2.1.2 Fixation-state capture

Following the identification of critical time windows of heightened attentional demand, the attentional state capture module categorizes fixation durations to characterize distinct cognitive processing levels. As informed by prior research, this study analyzes fixation states according to fixation duration thresholds: short fixations (<200 ms), medium fixations (200–500 ms), and long fixations (>500 ms) [32, 73]. These thresholds are employed because they map to attentional intensity, with shorter fixations reflecting quick, automatic processing of simple stimuli and longer fixations indicating complex cognitive effort or high cognitive load, as observed in tasks like reading and visual search[72–74]. This categorization enables precise quantification of attentional states, providing a foundation for analyzing cognitive dynamics within pre-alert periods.

5.2.1.3 Fixation-state transition feature set

The state transitions modeling module illustrates the construction of fixation-state transition feature set, depicting the fixation-state transition matrix, per-state frequency statistics, and state-sequence spatial patterns, as shown in Figure 5.2. First, the module extends the characterization of attentional states by constructing a fixation-state transition matrix to quantify the dynamic shifts between these attentional states. The transition matrix quantifies transition counts between short / medium / long fixation states within each time window, forming a 3×3 matrix yielding a 9-dimensional feature space. For states $s_i, s_j \in \{\text{short}, \text{medium}, \text{long}\}$, the transition count is defined as:

$$C(s_i \to s_j) = N(s_i \to s_j), \tag{5.1}$$

where $N(s_i \to s_j)$ is the number of transitions from state s_i to s_j. The resulting fixation-state transition counts include: $C_{\text{long}\to\text{long}}$, $C_{\text{long}\to\text{medium}}$, $C_{\text{long}\to\text{short}}$, $C_{\text{medium}\to\text{long}}$, $C_{\text{medium}\to\text{medium}}$, $C_{\text{medium}\to\text{short}}$, $C_{\text{short}\to\text{long}}$, $C_{\text{short}\to\text{medium}}$, and $C_{\text{short}\to\text{short}}$. To complement the transition matrix analysis, we compute per-state frequency statistics and state-sequence spatial patterns that characterize the overall distribution of attentional investment across different processing levels. Per-state fre-

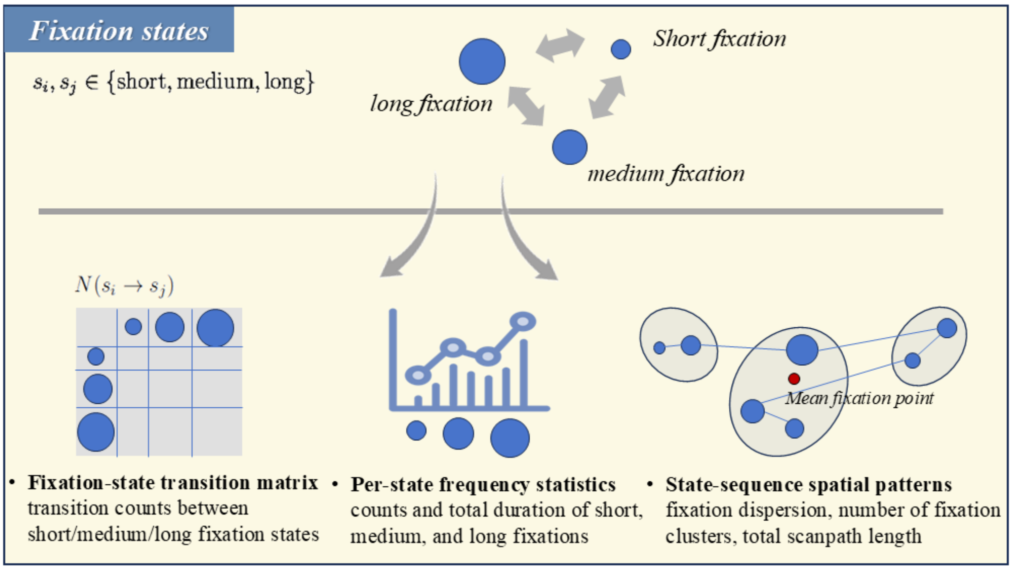


Figure 5.2: Conceptual diagram of fixation-state transition feature set, depicting the fixation-state transition matrix, per-state frequency statistics, and state-sequence spatial patterns.

quency statistics, including the respective counts and total counts of short, medium, and long fixations, and total fixation duration within each time window, provide a quantitative summary of attentional state distribution [71]. Additionally, state-sequence spatial patterns are characterized through three complementary metrics: total scanpath length (reflecting visual coverage), fixation dispersion (quantifying spatial distribution), and number of fixation clusters (indicating attentional focal points). These spatial features, inspired by prior work [74, 177], enhance the feature set by linking state transitions to spatial attention patterns, such as clustering of long fixations on complex regions, which is particularly relevant for tasks requiring strategic visual exploration.

Together, these features provide a comprehensive representation of attentional dynamics. To ensure robustness, features are standardized per participant using

StandardScaler to account for individual variability.

5.2.2 DF prediction based on L-PTT

The DF prediction module integrates a structured pipeline to enhance class-specifc probabilistic threshold tuning, as depicted in Figure 5.1. The process begins with data partitioning, where the dataset is initially split with 20% reserved as an independent test set to evaluate model performance. LLMs are employed to learn feature characteristics and derive probability threshold decisions from the training samples, which are subsequently applied to the test data. To ensure robust model evaluation, the remaining 80% of the data undergoes 5-fold cross-validation, further partitioned into training and validation subsets for iterative model tuning and assessment [187]. Additionally, to address class imbalance in the training data, SMOTE with $k = 5$ neighbors is applied to oversample the training folds, balancing rare DF classes. Two robust-performing models identified in prior work on DF recognition task were employed in this DF prediction task: random forest and Transformer-MLP [177]. Specifically, random forest utilizes ensemble learning with 100 trees, Gini criterion, max depth 20, while Transformer-MLP used PyTorch 1.9 with a 4-layer transformer encoder (8 attention heads, 512 units, dropout 0.1) followed by a 2-layer MLP (256 units, ReLU activation) to capture sequential fixation dependencies.

To address the challenges in specialized DF prediction, this study introduces L-PTT method, which refines classification decisions through class-specific thresholds. The process begins with computing the sample-specific predictive probability in each class using baseline models trained on the proposed feature set.

These probabilities reflect the model's confidence for each test instance across the four DF categories. Next, to reduce computational overhead, feature importance is evaluated globally via 5-fold cross-validation, with averaged scores (e.g., Gini impurity) ranking features. Only the top-ranked features (e.g., $k = 2, 3, 4, 5$) are retained, focusing on the most predictive ones to optimize efficiency. The selected features are then used to generate the probability distribution in each class and top-ranked features, capturing how feature values correlate with predicted probabilities across DF categories. This distribution is represented in two formats as input to the LLM: a 2D (two-dimensional) heatmap visualizing probability densities (e.g., using kernel density estimation for smooth interpolation) or a numerical table (e.g., CSV table of raw values). By comparing these formats, we evaluate which better enables the LLM to interpret feature-probability relationships for threshold calibration.

The core LLM-driven probabilistic threshold tuning step employs GPT-4-turbo via the OpenAI API to analyze the input representations and generate class-specific thresholds. For each class i ($i = 0, 1, 2, 3$; DF classes: $i = 1, 2$), the LLM-driven threshold is calculated as:

$$T_i = T_i^{init} + \Delta T_i(x) \tag{5.2}$$

where T^{init} is an initial threshold, and $\Delta T_i(x)$ is a feature-conditioned adjustment with x representing the input sample's feature vector. This adjustment captures the relationship between feature space and probability intensity, enabling context-aware threshold adaptation.

In addition, the LLM's workflow process and prompt design are detailed in

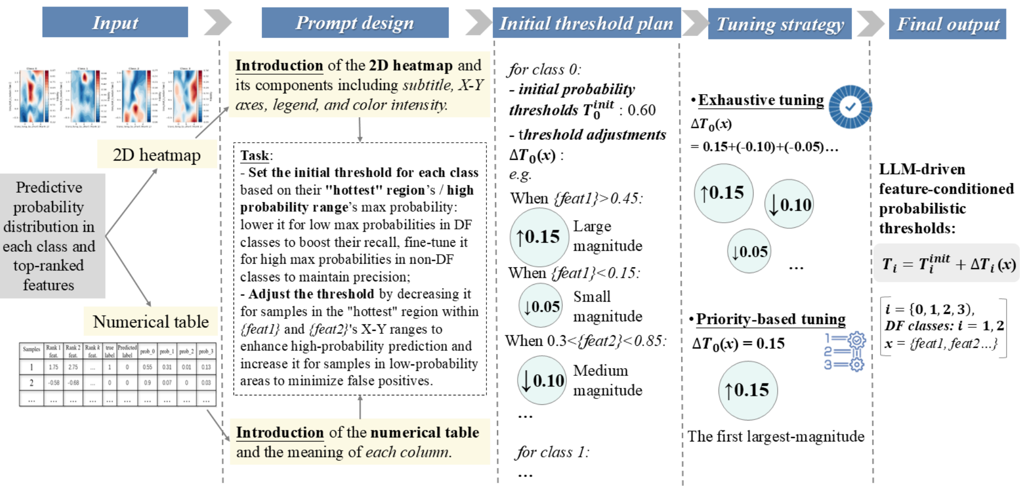


Figure 5.3: Workflow of LLM-driven probabilistic threshold tuning (L-PTT) method.

Figure 5.3, with probability distributions (as input) in 2D heatmap or numerical table formats. The prompt guides the LLM to set initial thresholds based on high-probability regions and adjust them dynamically—lowering for DF classes to enhance recall and refining for non-DF classes to preserve precision, while adapting thresholds based on feature-driven probability intensity to reduce false positives. The output comprises a Threshold Plan for each class, detailing initial thresholds, feature-conditioned adjustments (including conditions, actions, values, and magnitudes), and an explanatory basis for the adjustments. The generated thresholds are applied through two tuning strategies. Exhaustive tuning sequentially applies all LLM-proposed adjustments to optimize across feature combinations, ensuring comprehensive performance gains. Priority-based tuning selects only the largest-magnitude adjustment per sample (applying the first satisfied condition when equal), prioritizing efficiency while maintaining focus on impactful shifts.

Finally, sample-specific prediction intervenes in the classification decision as:

$$y = \arg\max_{i|P_i \geq T_i} P_i \tag{5.3}$$

where P_i is the predictve probability in each class. The LPT method filters probabilities that meet or exceed the threshold before selecting the maximum among qualified classes.

5.2.3 Performance Evaluation

Performance evaluation employs three key metrics: accuracy measuring overall detection correctness, class-wise recall assessing DF-category prediction capability, and class-wise F1-score balancing precision and recall - particularly crucial for handling class imbalance by penalizing scenarios where high recall might otherwise inflate false positives [20]. The proposed feature set is rigorously compared against two established baselines: (1) a classical feature set comprising 17 traditional eye-tracking metrics, where the number of features was deliberately matched to our method's dimensionality while maintaining comprehensive coverage of fundamental attentional indicators. The metrics include average/total fixation duration, fixation duration, fixation frequency, fixation count, saccade count, saccade frequency, average/maximum saccade velocity, average/maximum blink rate, and mean pupil diameter (left and right), fixation-saccade ratio, average/maximum saccade magnitude, as extensively documented in previous research on eye movement analysis [16, 25, 75, 162]. and (2) a SOTA fixation coordinate-sensitive feature set that demonstrated superior performance in prior work through its 12-dimensional spatial-temporal characterization of fixation patterns [177].

The L-PTT method undergoes comprehensive benchmarking against established techniques, including probability calibration approaches [93] and threshold optimization methods (ROC/PR curve-based tuning [183] and grid search [182]) for post-processing, as well as data pre-processing, the standard SMOTE and ADASYN to evaluate their impact on class balance [87]. Additionally, this study also compares class balancing strategies such as Class-Weighted learning, BalancedSampling, and EasyEnsemble methods [88], and recent algorithmic optimization techniques [91, 180], including BOHB and BO-TPE, assessing their effectiveness in hyperparameter tuning for the ATC-specific dataset.

5.3 Results

5.3.1 Evaluation of fixation-state transition feature set

This section evaluates the proposed fixation-state transition feature set against traditional and fixation coordinate-sensitive feature sets across three pre-alert periods (i.e., the Focused (0 to -5s), Moderate (0 to -9s), and Extended (0 to -14s) windows) using random forest and Transformer-MLP models, as shown in Tables 5.1 and 5.2. The analysis prioritizes overall accuracy for prediction reliability and recall/F1-score for minority DF types, critical for accurate DF prediction.

For the random forest model (see Table 5.1), the proposed feature set consistently outperforms Trad. and FCS across all time ranges, achieving the highest accuracy (0.78) and minority class metrics in the Extended period (DF-1 recall: 0.35, F1-score: 0.49; DF-2 recall: 0.31, F1-score: 0.45). The Extended time range yields the best overall performance, with the proposed feature set improving DF-1

Table 5.1: Performance metrics comparison of random forest across different time ranges and feature sets. Here, recall rate and F1-score focus on two DF types.

Time ranges	Feat. set	Acc.	Recall rate				F1-score			
			NDF-1	DF-1	DF-2	NDF-2	NDF-1	DF-1	DF-2	NDF-2
Focused	Trad.	**0.69**	1.00	0.00	0.00	0.00	0.83	0.00	0.00	0.00
	FCS	0.65	0.93	0.00	0.00	0.00	0.79	0.00	0.00	0.00
	FTFS	0.67	0.97	0.00	0.00	0.00	0.80	0.00	0.00	0.00
Moderate	Trad.	0.70	1.00	0.06	0.05	0.02	0.82	0.11	0.09	0.03
	FCS	**0.73**	0.95	0.23	**0.29**	0.19	0.83	0.35	**0.39**	0.29
	FTFS	**0.73**	0.92	**0.42**	**0.29**	0.18	0.83	**0.51**	**0.39**	0.26
Extended	Trad.	0.71	1.00	0.07	0.05	0.05	0.83	0.13	0.09	0.11
	FCS	0.73	0.97	0.14	0.12	0.28	0.84	0.25	0.21	0.40
	FTFS	**0.78**	0.97	**0.35**	**0.31**	0.36	0.86	**0.49**	**0.45**	0.49

Note:
- NDF denotes non-DF types, DF denotes DF types; Trad., FCS, FTFS denote traditional feature set, fixation coordinate-sensitive feature set, fixation-state transition feature set (proposed), respectively.
- Bold values indicate the best performance of each metric class in DF types.

Table 5.2: Performance metrics comparison of Transformer-MLP across different time ranges and feature sets. Here, recall rate and F1-score focus on two DF types.

Time Range	Feat. Set	Acc.	Recall Rate				F1-score			
			NDF-1	DF-1	DF-2	NDF-2	NDF-1	DF-1	DF-2	NDF-2
Focused	Trad.	0.35	0.44	0.00	**0.25**	0.25	0.55	0.00	**0.13**	0.16
	FCS	0.41	0.56	**0.20**	0.00	0.00	0.61	**0.13**	0.00	0.00
	FTFS	**0.61**	0.81	**0.20**	0.00	0.17	0.76	**0.22**	0.00	0.19
Moderate	Trad.	0.57	0.71	**0.29**	0.10	0.30	0.72	**0.29**	0.11	0.27
	FCS	0.54	0.64	**0.29**	**0.33**	0.30	0.70	0.25	**0.21**	0.29
	FTFS	**0.58**	0.73	0.27	**0.33**	0.23	0.73	0.26	**0.34**	0.22
Extended	Trad.	0.60	0.77	0.17	0.17	0.32	0.75	0.19	0.20	0.30
	FCS	0.58	0.73	0.22	0.19	0.32	0.73	0.23	0.16	0.32
	FTFS	**0.62**	0.75	**0.35**	**0.21**	0.32	0.75	**0.35**	**0.25**	0.29

Note:
- NDF denotes non-DF types, DF denotes DF types; Trad., FCS, FTFS denote traditional feature set, fixation coordinate-sensitive feature set, fixation-state transition feature set (proposed), respectively.
- Bold values indicate the best performance of each metric class in DF types.

recall by 5 $times$ (0.07 to 0.35) and DF-2 recall by 6 $times$ (0.05 to 0.31) compared to Trad. in the same period. In contrast, the Focused period yields zero recall for DF-1/DF-2 across all feature sets, as the single 5-second window and small test set (e.g., 4 DF-2 samples) limit feature discriminability. The Moderate period shows moderate gains (the proposed feature set: DF-1 recall 0.42, DF-2 recall 0.29), but Extended's larger sample size and temporal context improve overall performance.

By comparison, the Transformer-MLP model (see Table 5.2) exhibits less effective overall performance than random forest, with the proposed feature set still achieving the highest accuracy (0.62 in Extended) and minority class metrics (DF-1 recall: 0.35, F1-score: 0.35; DF-2 recall: 0.21, F1-score: 0.25). Compared to Trad. and FCS, the proposed feature set improves DF-1 recall by 2 $times$ (0.17 to 0.35) and DF-2 F1-score by 25% (0.20 to 0.25) in the Extended period. However, Transformer-MLP also struggles in the Focused period (DF-1 recall: 0.20, DF-2 recall: 0.00 for the proposed feature set), indicating sensitivity to limited temporal data. The Extended period again proves optimal, leveraging the proposed feature set's transition-focused features to enhance minority class detection, though random forest's superior robustness suggests better suitability for imbalanced datasets.

Given random forest's superior performance and the proposed feature set's consistent effectiveness, particularly in the Extended time range, subsequent analyses adopt the random forest model with the Extended period and the proposed feature set to further optimize DF prediction. This choice ensures robust minority class performance while maintaining conciseness in evaluating the L-PTT method's adaptive threshold strategies. Under this optimal setting, the top-5 features by global importance are: $C_{\text{long}\to\text{short}}$ (0.10), $C_{\text{short}\to\text{long}}$ (0.07), short fixation count (0.07), fixation dispersion (0.07), and $C_{\text{short}\rightarrow\text{medium}}$ (0.07). These features, capturing frequent attentional shifts and spatial variability, are critical for distinguishing DF-1 and DF-2 by highlighting dynamic attention switching patterns.

5.3.2 Evaluation of DF prediction based on L-PTT

Table 5.3: Performance evaluation of L-PTT methods in different combinations of top-k features, input forms and tuning strategies.

Top-k	Methods	Acc.	Recall Rate				F1-score			
			NDF-1	DF-1	DF-2	NDF-2	NDF-1	DF-1	DF-2	NDF-2
2	H-Pri	0.72	0.71	**0.75**	**0.64**	0.78	0.80	0.57	**0.61**	0.62
	H-Exh	0.76	0.79	0.70	0.62	0.73	0.84	**0.61**	0.55	0.65
	A-Pri	0.77	0.96	0.36	0.36	0.38	0.86	0.47	0.49	0.51
	A-Exh	**0.78**	0.94	0.48	0.38	0.38	0.86	0.55	0.52	0.51
3	H-Pri	**0.80**	0.89	0.54	**0.62**	0.65	0.87	0.57	**0.62**	0.67
	H-Exh	0.78	0.89	**0.64**	0.57	0.45	0.87	**0.58**	0.59	0.57
	A-Pri	0.78	0.97	0.31	0.31	0.38	0.86	0.44	0.45	0.51
	A-Exh	0.78	0.97	0.31	0.31	0.38	0.86	0.44	0.45	0.51
4	H-Pri	**0.79**	0.86	**0.55**	**0.64**	0.67	0.86	**0.58**	**0.61**	0.66
	H-Exh	0.78	0.95	0.37	0.33	0.46	0.86	0.49	0.47	0.56
	A-Pri	0.77	0.96	0.38	0.31	0.38	0.86	0.49	0.45	0.51
	A-Exh	0.77	0.97	0.31	0.29	0.35	0.86	0.44	0.42	0.49
5	H-Pri	0.76	0.83	**0.57**	**0.67**	0.59	0.84	**0.58**	0.45	0.64
	H-Exh	**0.78**	0.97	0.32	0.36	0.38	0.86	0.45	**0.48**	0.51
	A-Pri	0.77	0.96	0.38	0.31	0.38	0.86	0.49	0.45	0.51
	A-Exh	0.75	0.88	0.35	0.31	0.59	0.84	0.46	0.45	0.56

Method Key: H-Pri = Heatmap-Priority, H-Exh = Heatmap-Exhaustive
A-Pri = Array-Priority, A-Exh = Array-Exhaustive

Note: Bold values indicate the best performance in each metric class.

Building on the Section 5.3.1 findings, where random forest model with the extended time range and FTFS achieved the highest performance, limitations in recall and F1-score for critical DF types, DF-1 and DF-2, necessitate further optimization. This section introduces the L-PTT approach, which employs adaptive threshold tuning to enhance recall and F1-score for these minority DF types. Table 5.3

evaluates four L-PTT methods—Heatmap-Priority (H-Pri), Heatmap-Exhaustive (H-Exh), Array-Priority (A-Pri), and Array-Exhaustive (A-Exh)—using two input forms (2D heatmaps and numerical arrays) and two threshold tuning strategies (exhaustive and priority-based). The table reports accuracy, recall, and F1-score for four classes, with DF-1 and DF-2 being critical for improved recall and F1-score due to their sparse representation. The H-Pri method demonstrates robust performance across all top-k feature sets, achieving the highest accuracy (0.80 for top-3, 0.79 for top-4, 0.76 for top-5) and strong recall for DF-1 (0.75, 0.55, 0.57) and DF-2 (0.64, 0.64, 0.67). H-Exh also performs competitively, particularly for top-2 (0.76 accuracy, 0.70 recall for DF-1) and top-3 (0.64 recall for DF-1), but its minority class performance declines with larger feature sets (e.g., 0.32 recall for DF-1 in top-5). In contrast, A-Pri and A-Exh consistently favor the majority class ED (recall up to 0.97), but their performance on DF-1 and DF-2 (recall 0.31–0.38) indicates a bias due to data imbalance. Increasing feature set size from top-2 to top-5 does not always improve performance, as seen in A-Exh's accuracy drop (0.78 to 0.75), suggesting potential overfitting with larger feature sets.

Given H-Pri's superior accuracy (0.80) with the top-3 feature set, 2D heatmaps (Figure 6.2 in Appendix IV) were provided to visualize the joint probability distributions of pairwise combinations among the top-3 features (i.e., $C_{\text{long}\to\text{short}}$, $C_{\text{short}\to\text{long}}$, short fixation count) in the training set. These heatmaps, serving as inputs to the L-PTT model, reveal lower probability densities for DF-1 and DF-2 compared to ED, highlighting the challenge of imbalanced data in distinguishing critical DF types. The focus on top-3 feature pairs is justified by H-Pri's optimal balance of high accuracy and improved minority class performance, providing insights into feature interactions that enhance DF prediction.

Table 5.4: Comparative performance of L-PTT method against SOTA DF prediction methods addressing imbalance issues.

Method	Acc.	Recall Rate				F1-score			
		NDF-1	DF-1	DF-2	NDF-2	NDF-1	DF-1	DF-2	NDF-2
SOTA Methods									
SMOTE-only	0.78	0.97	0.31	0.31	0.38	0.86	0.44	0.45	0.51
Class-Weighted	0.78	0.97	0.31	0.31	0.38	0.86	0.44	0.45	0.51
Balanced Sampling	0.57	0.50	**0.74**	**0.83**	0.68	0.65	0.52	0.41	0.50
EasyEnsemble	0.37	0.34	0.46	0.52	0.41	0.47	0.30	0.20	0.30
BOHB	**0.80**	0.90	0.53	0.55	0.61	0.87	0.56	0.60	0.66
BO-TPE	<u>0.79</u>	0.89	0.54	0.55	0.63	0.87	<u>0.58</u>	0.57	0.67
PR-Threshold	0.78	0.97	0.31	0.31	0.38	0.86	0.44	0.45	0.51
ROC-Threshold	<u>0.79</u>	0.94	0.41	0.48	0.52	0.87	0.52	0.53	0.61
Grid Search	0.78	0.94	0.38	0.38	0.49	0.86	0.49	0.52	0.59
Probability-Calibrated	0.78	0.95	0.38	0.38	0.40	0.86	0.49	0.51	0.52
Proposed Methods									
H-Exh (ADASYN)	0.78	0.95	0.35	0.60	0.38	0.86	0.48	0.60	0.51
H-Exh (SMOTE)	0.78	0.89	<u>0.64</u>	0.57	0.45	0.87	<u>0.58</u>	0.59	0.57
H-Pri (SMOTE)	**0.80**	0.89	0.54	0.62	0.65	0.87	0.57	<u>0.62</u>	0.67
H-Pri (ADASYN)	**0.80**	0.89	0.53	<u>0.64</u>	0.65	0.87	**0.59**	**0.64**	0.67

Note:
- NDF denotes non-DF types, DF denotes DF types; Trad., FCS, FTFS denote traditional feature set, fixation coordinate-sensitive feature set, fixation-state transition feature set (proposed), respectively.
- Bold values indicate the best performance of each metric class in DF types.

5.3.3 Comparative analysis with SOTA methods

The previous Section 5.3.2 analysis established the superiority of the top-3 features with Heatmap input. Hence, this study evaluates the L-PTT method using this configuration with two application strategies—Priority (H-Pri) and Exhaustive (H-Exh), each with SMOTE or ADASYN preprocessing, against state-of-the-art (SOTA) methods for DF prediction with class imbalance issue, as detailed in Table 5.4. L-PTT's H-Pri (SMOTE) and H-Pri (ADASYN) achieve a top accuracy of 0.80, matching BOHB, while surpassing it in DF-2 recall (0.62–0.64 vs. 0.55), and F1-scores. Although Balanced Sampling achieves high DF-1 (0.74) and

DF-2 (0.83) recall, its low accuracy (0.57) risks excessive false positives, compromising operational reliability. In contrast, H-Exh (SMOTE) excels in DF-1 recall (0.64), second only to Balanced Sampling's 0.74, and H-Pri (ADASYN) achieves superior F1-scores (0.59 DF-1, 0.64 DF-2) alongside high accuracy and achieves the second-highest DF-2 recall (0.64, behind Balanced Sampling's 0.83). Additionally, unlike PR-Threshold, ROC-Threshold, and GridSearch, which use static thresholds and underperform (0.31–0.48 DF-1/DF-2 recall, 0.44–0.53 F1-scores), L-PTT's adaptive feature-conditioned threshold enhances minority DF prediction. Other SOTA methods, such as SMOTE-only, Class-Weighted, EasyEnsemble, BO-TPE, and Probability-Calibrated, show limited performance (0.31–0.54 DF-1 recall, 0.20–0.58 F1-scores), with EasyEnsemble notably weak (0.37 accuracy). In summary, L-PTT's H-Pri (ADASYN) ensures reliable DF prediction without compromising overall performance.

5.4 Discussion

5.4.1 Key findings and methodological implications

The results demonstrate that the proposed feature set significantly outperforms traditional feature sets, particularly in the Extended time window where it achieves 0.78 accuracy (vs. 0.71-0.73 baselines) and high improvement in DF-type recall. This superiority stems from the proposed feature set's unique capacity to encode temporal attentional transitions (e.g., long-to-short fixation sequences), which conventional spatial features or static metrics fail to capture. The random forest model's consistent advantage over Transformer-MLP (0.78 vs 0.62 accuracy) suggests ensemble methods better handle the imbalanced DF data distribution, likely through their inherent bootstrap aggregation, mitigating overfitting to majority classes. Notably, the Extended period's performance dominance reveals a critical trade-off: while longer pre-alert windows (14s) provide richer contextual cues for detection, they may delay response times in operational settings - a tension requiring careful calibration in real-world deployment. The top-performing features ($C_{\text{long}\rightarrow\text{short}}$ (0.10), $C_{\text{short}\rightarrow\text{long}}$ (0.07), short fixation count) empirically validate cognitive theories that abrupt attentional shifts precede DF, offering neuroscientific grounding for our approach. Moreover, the results from the L-PTT Optimization approach reveal that the top three features provide a more effective framework for large language model (LLM) learning compared to configurations incorporating the top two, four, or five features, suggesting a refined feature set enhances model comprehension and performance [16]. In addition, LLM learning from 2D heatmaps demonstrates superior efficacy over traditional numeric structure-based learning, indicating that simplified visual representations facilitate the model's ability to discern and prioritize key patterns effectively.

5.4.2 Practical applications and future directions

L-PTT's adaptive threshold tuning proves particularly valuable for safety-critical ATC operations, where its 0.80 accuracy and 0.53-0.64 minority recall outperform static methods while avoiding Balanced Sampling's false positive risks (57% accuracy). The success of heatmap inputs (vs. numerical arrays) suggests visual feature representations better preserve spatial-temporal relationships crucial for DF prediction, aligning with recent vision transformer advancements [29, 96]. However, three limitations warrant future work: The pre-alert period's time ranges (Focused, Moderate, Extended) and sampling methods require finer-grained exploration to optimize temporal granularity, as coarse intervals may obscure subtle attentional shifts critical for DF detection. Additionally, L-PTT's LLM-based threshold tuning could benefit from enhanced prompt engineering to improve adaptability to diverse feature interactions visualized in heatmaps. Finally, the real-time application of LLMs in ATC systems faces computational constraints, limiting immediate deployment despite robust performance. These challenges suggest future research into refined temporal sampling and efficient LLM integration.

5.5 Concluding remarks

This study introduces a methodological reference for proactive DF prediction in ATC by integrating fixation-state transition feature set and L-PTT. The results demonstrate that the proposed feature set effectively captures pre-alert attentional dynamics, achieving 0.78 accuracy (vs. 0.71–0.73 baselines) and significantly enhancing recall for critical DF types (DF-1: 0.35 → 0.53–0.64; DF-2: 0.31 → 0.59–0.64). L-PTT further optimizes performance through adaptive thresholding, outperforming static methods (0.80 accuracy) while maintaining robustness. The superiority of 2D heatmap inputs and priority-based tuning highlights the importance of spatially aware feature representation and dynamic decision-making. By providing practical guidance for reliable, human-centric DF prediction, this work enhances ATC safety and sets a foundation for LLM-driven adaptive systems in safety-critical domains.

Chapter 6

Conclusion

This chapter concludes the thesis by summarizing the key findings and main contributions in Section 6.1, and discussing the study's limitations along with proposed directions for future research in Section 6.2.

6.1 Contributions

This thesis significantly advances the theoretical classification, recognition, and prediction of visual DF in ATC through the integration of eye-tracking analytics, cognitive modeling, and LLM. The research establishes novel theoretical frameworks, develops computationally informed methodologies, and provides actionable insights for enhancing safety and operational efficiency in ATC systems. The key contributions are summarised as follows:

1. **Situation Awareness-Based Theoretical Foundation for DF Classification**

- Proposed a novel gaze-behavioral DF classification grounded in Endsley's situation awareness theory, systematically categorizing DF and linking diverse DF types to specific eye-tracking indicators.
- Revealed how warning frequency modulates operators' attentional patterns and DF types, offering theoretical insights into the interaction between external stimuli and visual-cognitive behaviors in dynamic environments.
- Provided a scientific basis for the design of adaptive warning systems, supporting dynamic frequency adjustment and cognitive-state-aware interventions.

2. **Hierarchical and Peripheral Vision-Aware DF Recognition Framework**

- Developed an ATC-specific hierarchical recognition framework integrating peripheral vision tracking and multi-domain gaze feature extraction, significantly improving DF classification accuracy under realistic operational conditions.
- Introduced an Adaptive Symbolic Alert Detection method for automatic identification and annotation of ATC alerts, enhancing the efficiency and objectivity of alert region localization.
- Leveraged GPT-4V for evaluating controllers' visual engagement with alerts, reducing reliance on manual annotation and enabling scalable, intelligent alert attention assessment.

3. **Fixation State-Driven and LLM-Enhanced DF Predictive Modeling**

- Designed a novel fixation-state transition feature set to capture pre-alert temporal attention patterns, enabling early and accurate prediction of DF.
- Proposed an LLM-driven probabilistic threshold tuning method that adaptively sets classification thresholds based on feature-probability heatmaps, improving robustness in class-imbalanced scenarios.
- Demonstrated the advantage of 2D visual heatmaps over numerical data in facilitating LLM interpretability, providing a new pathway for explainable AI in safety-critical human-machine systems.

In summary, this thesis contributes to both the theoretical and practical dimensions of human factors in ATC by introducing scalable, interpretable, and fixation-aware frameworks for DF analysis. The methodologies developed are also transferable to other safety-critical monitoring domains such as autonomous driving, medical supervision, and power grid management.

6.2 Limitations and Future Research

While this thesis provides substantial contributions to DF research, several limitations remain that offer avenues for future investigation:

1. *Experimental Scope and Scenario Complexity*: Current studies employed simplified ATC tasks with visual-only alerts. Future work should incorporate multi-modal alerts (e.g., auditory cues) and overlapping warning scenarios to better approximate real-world conditions.

2. *Sample Size and Participant Diversity*: Constraints in data collection within operational ATC settings resulted in a limited number of participants. Future studies should include more controllers with varying expertise levels to improve generalizability.

3. *Temporal and Computational Granularity*: The predefined pre-alert time intervals and high computational demands of LLMs pose challenges for real-time deployment. Future research should explore finer-grained temporal segmentation and efficient model compression or hardware acceleration strategies.

4. *Integration of Multi-Modal Data*: This thesis focused primarily on eye-tracking data. Incorporating EEG and other physiological signals could provide a more comprehensive understanding of cognitive states during DF episodes.

5. *Context Adaptation of LLM Prompts*: The zero-shot prompts used for alert evaluation and threshold tuning showed limitations in task-specific accuracy. Fine-tuning LLMs on ATC-specific corpora may enhance performance and contextual alignment.

Future research will focus on extending the proposed frameworks to more complex and dynamic environments, integrating multi-modal data sources, enhancing computational efficiency, and refining LLM adaptations for real-time, safety-critical applications.

References

[1] Xu, R., Luo, F., Zhou, F., Chen, G. “Risk early warning indicators for air traffic controllers’ unsafe acts”. In: *Human Factors and Ergonomics in Manufacturing & Service Industries* 32.3 (2022), pp. 286–300.

[2] Varakin, D. A., Levin, D. T., Fidler, R. “Unseen and unaware: Implications of recent research on failures of visual awareness for human-computer interface design”. In: *Human–Computer Interaction* 19.4 (2004), pp. 389–422.

[3] Shorrock, S. T. “Errors of perception in air traffic control”. In: *Safety science* 45.8 (2007), pp. 890–904.

[4] Giraudet, L., Imbert, J.-P., Tremblay, S., Causse, M. “High rate of inattentional deafness in simulated air traffic control tasks”. In: *Procedia Manufacturing* 3 (2015), pp. 5169–5175.

[5] Kennedy, K. D., Stephens, C. L., Williams, R. A., Schutte, P. C. “Automation and inattentional blindness in a simulated flight task”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 58. 1. SAGE Publications Sage CA: Los Angeles, CA. 2014, pp. 2058–2062.

[6] Nikkei, *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`. 2024.

[7] Wang, Y., Wu, Y., Chen, C., Wu, B., Ma, S., Wang, D., Li, H., Yang, Z. "Inattentional blindness in augmented reality head-up display-assisted driving". In: *International Journal of Human–Computer Interaction* 38.9 (2022), pp. 837–850.

[8] White, A., O'Hare, D. "In plane sight: Inattentional blindness affects visual detection of external targets in simulated flight". In: *Applied Ergonomics* 98 (2022), p. 103578.

[9] Simons, D. J., Chabris, C. F. "Gorillas in our midst: Sustained inattentional blindness for dynamic events". In: *perception* 28.9 (1999), pp. 1059–1074.

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

[11] Mack, A., Rock, I. "Inattentional blindness: Perception without attention". In: *Visual attention* 8.01 (1998).

[12] Mack, A. "Inattentional blindness: Looking without seeing". In: *Current directions in psychological science* 12.5 (2003), pp. 180–184.

[13] Li, Q., Yiu, C. Y., Simon, C., Ng, K. K. "Situational awareness and flight approach phase event recognition based on psychophysiological

measurements”. In: *2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. IEEE. 2021, pp. 1308–1312.

[14] Thomas, L. C., Wickens, C. D. “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*. Vol. 45. 4. SAGE Publications Sage CA: Los Angeles, CA. 2001, pp. 336–340.

[15] White, C. B. “Looked but failed to see errors”. In: (2008).

[16] Mengtao, L., Fan, L., Gangyan, X., Su, H. “Leveraging eye-tracking technologies to promote aviation safety-a review of key aspects, challenges, and future perspectives”. In: *Safety science* 168 (2023), p. 106295.

[17] Li, F., Chen, C.-H., Xu, G., Khoo, L.-P. “Hierarchical eye-tracking data analytics for human fatigue detection at a traffic control center”. In: *IEEE Transactions on Human-Machine Systems* 50.5 (2020), pp. 465–474.

[18] Li, F. “Data-driven, context-aware human fatigue management in traffic control centers”. In: (2020).

[19] Ruskin, K. J., Corvin, C., Rice, S., Richards, G., Winter, S. R., Ruskin, A. C. “Alarms, alerts, and warnings in air traffic control: an analysis of reports from the Aviation Safety Reporting System”. In: *Transportation research interdisciplinary perspectives* 12 (2021), p. 100502.

[20] Lyu, M. “Visual attention-based pilots’ status monitoring and performance enhancement towards human-in-the-loop automation”. In: (2025).

[21] Bruder, C., Hasse, C. “What the eyes reveal: Investigating the detection of automation failures”. In: *Applied ergonomics* 82 (2020), p. 102967.

[22] Friedman-Berg, F., Allendoerfer, K., Pai, S. “Nuisance alerts in operational ATC environments: Classification and frequencies”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 52. 1. SAGE Publications Sage CA: Los Angeles, CA. 2008, pp. 104–108.

[23] Dreyfus, H. L., Dreyfus, S. E. “Peripheral vision: Expertise in real world contexts”. In: *Organization studies* 26.5 (2005), pp. 779–792.

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

[25] **Li, Z.**, Li, R., Yuan, L., Cui, J., Li, F. “A benchmarking framework for eye-tracking-based vigilance prediction of vessel traffic controllers”. In: *Engineering Applications of Artificial Intelligence* 129 (2024), p. 107660.

[26] Li, F., Chen, C.-H., Xu, G., Khoo, L. P., Liu, Y. “Proactive mental fatigue detection of traffic control operators using bagged trees and gaze-bin analysis”. In: *Advanced Engineering Informatics* 42 (2019), p. 100987.

[27] Xu, R., Luo, F. “Risk prediction and early warning for air traffic controllers’ unsafe acts using association rule mining and random forest”. In: *Safety science* 135 (2021), p. 105125.

[28] Zhou, F., Yang, X. J., De Winter, J. C. “Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving”. In: *IEEE transactions on intelligent transportation systems* 23.3 (2021), pp. 2284–2295.

[29] Li, F., Han, S., Lee, C.-H., Feng, S., Jiang, Z., Sun, Z. "A new era in human factors engineering: A survey of the applications and prospects of large multimodal models". In: *International Journal of Human–Computer Interaction* (2025), pp. 1–14.

[30] Endsley, M. R. "Toward a theory of situation awareness in dynamic systems". In: *Human factors* 37.1 (1995), pp. 32–64.

[31] Wu, Z., Zhao, L., Liu, G., Chai, J., Huang, J., Ai, X. "The Effect of AR-HUD Takeover Assistance Types on Driver Situation Awareness in Highly Automated Driving: A 360-Degree Panorama Experiment". In: *International Journal of Human–Computer Interaction* (2023), pp. 1–18.

[32] Negi, S., Mitra, R. "Fixation duration and the learning process: An eye tracking study with subtitled videos". In: *Journal of Eye Movement Research* 13.6 (2020), pp. 10–16910.

[33] Yang, J., Liang, N., Pitts, B. J., Prakah-Asante, K., Curry, R., Yu, D. "An eye-fixation related electroencephalography technique for predicting situation awareness: implications for driver state monitoring systems". In: *Human factors* 66.8 (2024), pp. 2138–2153.

[34] Sheng, V. S., Ling, C. X. "Thresholding for making classifiers cost-sensitive". In: *Aaai*. Vol. 6. 2006, pp. 476–481.

[35] Imbert, J.-P., Hodgetts, H. M., Parise, R., Vachon, F., Dehais, F., Tremblay, S. "Attentional costs and failures in air traffic control notifications". In: *Ergonomics* 57.12 (2014), pp. 1817–1832.

[36] Ruskin, K. J., Rice, S., Corvin, C. "ATC Signaling Systems: A Review of the Literature on Alarms, Alerts, and Warnings". In: (2020).

[37] Norman, D. A. “Categorization of action slips.” In: *Psychological review* 88.1 (1981), p. 1.

[38] Rensink, R. A., O’regan, J. K., Clark, J. J. “To see or not to see: The need for attention to perceive changes in scenes”. In: *Psychological science* 8.5 (1997), pp. 368–373.

[39] Migliorini, Y., Imbert, J.-P., Roy, R. N., Lafont, A., Dehais, F. “Degraded states of engagement in air traffic control”. In: *Safety* 8.1 (2022), p. 19.

[40] Wickens, C. D., Alexander, A. L. “Attentional tunneling and task management in synthetic vision displays”. In: *The international journal of aviation psychology* 19.2 (2009), pp. 182–199.

[41] Green, D. M., Swets, J. A. *Signal detection theory and psychophysics*. Vol. 1. Wiley New York, 1966.

[42] Atick, J. J., Redlich, A. N. “Towards a theory of early visual processing”. In: *Neural computation* 2.3 (1990), pp. 308–320.

[43] Jones, D. G., Endsley, M. R. “Sources of situation awareness errors in aviation”. In: *Aviation, Space and Environmental Medicine* 67.6 (1996), pp. 507–512.

[44] Zhaoping, L., Dayan, P. “Pre-attentive visual selection”. In: *Neural Networks* 19.9 (2006), pp. 1437–1439.

[45] Pammer, K., Blink, C. “Attentional differences in driving judgments for country and city scenes: Semantic congruency in inattentional blindness”. In: *Accident Analysis & Prevention* 50 (2013), pp. 955–963.

[46] Lavie, N. “Perceptual load as a necessary condition for selective attention.” In: *Journal of Experimental Psychology: Human perception and performance* 21.3 (1995), p. 451.

[47] Eayrs, J., Lavie, N. “Establishing individual differences in perceptual capacity.” In: *Journal of Experimental Psychology: Human Perception and Performance* 44.8 (2018), p. 1240.

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

[49] Vance, A., Kirwan, B., Bjornn, D., Jenkins, J., Anderson, B. B. “What do we really know about how habituation to warnings occurs over time? A longitudinal fMRI study of habituation and polymorphic warnings”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2017, pp. 2215–2227.

[50] Kennedy, K. D., Stephens, C. L., Williams, R. A., Schutte, P. C. “Repeated induction of inattentional blindness in a simulated aviation environment”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 61. 1. SAGE Publications Sage CA: Los Angeles, CA. 2017, pp. 1959–1963.

[51] Pappas, J. M., Fishel, S. R., Moss, J. D., Hicks, J. M., Leech, T. D. “An eye-tracking approach to inattentional blindness”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 49. 17. SAGE Publications Sage CA: Los Angeles, CA. 2005, pp. 1658–1662.

[52] Beanland, V., Pammer, K. “Looking without seeing or seeing without looking? Eye movements in sustained inattentional blindness”. In: *Vision research* 50.10 (2010), pp. 977–988.

[53] Ranti, C., Jones, W., Klin, A., Shultz, S. “Blink rate patterns provide a reliable measure of individual engagement with scene content”. In: *Scientific reports* 10.1 (2020), p. 8267.

[54] Li, Q., Ng, K. K., Yiu, C. Y., Yuan, X., So, C. K., Ho, C. C. “Securing air transportation safety through identifying pilot’s risky VFR flying behaviours: An EEG-based neurophysiological modelling using machine learning algorithms”. In: *Reliability Engineering & System Safety* 238 (2023), p. 109449.

[55] Li, Q. “Enhancing flight operations safety: a neurophysiological approach to mental state prediction in commercial single-pilot operations”. In: (2024).

[56] Chen, Y.-T., Lee, H.-H., Shih, C.-Y., Chen, Z.-L., Beh, W.-K., Yeh, S.-L., Wu, A.-Y. “An effective entropy-assisted mind-wandering detection system using EEG signals of MM-SART database”. In: *IEEE Journal of Biomedical and Health Informatics* 26.8 (2022), pp. 3649–3660.

[57] Papastavrou, J. D., Lehto, M. R. “Improving the effectiveness of warnings by increasing the appropriateness of their information content: some hypotheses about human compliance”. In: *Safety science* 21.3 (1996), pp. 175–189.

[58] Kearney, P., Li, W.-C., Lin, J. J. “The impact of alerting design on air traffic controllers’ response to conflict detection and resolution”. In: *International journal of industrial ergonomics* 56 (2016), pp. 51–58.

[59] Kearney, P., Li, W.-C., Yu, C.-S., Braithwaite, G. “The impact of alerting designs on air traffic controller’s eye movement patterns and situation awareness”. In: *Ergonomics* 62.2 (2019), pp. 305–318.

[60] Zhang, M., Ma, S., Xu, R., Chen, T., Ding, Y., Luo, X. “Evaluating the impact of proactive warning systems on worker safety performance: An immersive virtual reality study”. In: *Safety Science* 186 (2025), p. 106774.

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

[62] Ruscio, D., Ciceri, M. R., Biassoni, F. “How does a collision warning system shape driver’s brake response time? The influence of expectancy and automation complacency on real-life emergency braking”. In: *Accident Analysis & Prevention* 77 (2015), pp. 72–81.

[63] Dixon, S. R., Wickens, C. D., McCarley, J. S. “On the independence of compliance and reliance: Are automation false alarms worse than misses?” In: *Human factors* 49.4 (2007), pp. 564–572.

[64] Chi, C.-F., Lin, Y.-H., Lan, W.-S. “Measurement of information processing load and visual load on a dynamic information processing task”. In: *Behaviour & Information Technology* 22.5 (2003), pp. 365–374.

[65] Deng, T., Yan, H., Qin, L., Ngo, T., Manjunath, B. “How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks”. In: *IEEE Transactions on Intelligent Transportation Systems* 21.5 (2019), pp. 2146–2154.

[66] Krueger, E., Schneider, A., Sawyer, B. D., Chavaillaz, A., Sonderegger, A., Groner, R., Hancock, P. “Microsaccades distinguish looking from seeing”. In: *Journal of Eye Movement Research* 12.6 (2019), pp. 10–16910.

[67] Moacdieh, N. M., Dibo, M., Halabi, Z., Antoun, J. “Eye tracking to evaluate the effectiveness of electronic medical record training”. In: *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*. 2023, pp. 1–7.

[68] Baccour, M. H., Driewer, F., Schäck, T., Kasneci, E. “Comparative analysis of vehicle-based and driver-based features for driver drowsiness monitoring by support vector machines”. In: *IEEE transactions on intelligent transportation systems* 23.12 (2022), pp. 23164–23178.

[69] Krejtz, K., Szczecinski, P., Pawlowska, A., Rutkowska-Siuda, D., Wisiecka, K., Milczarski, P., Hlobaz, A., Duchowski, A. T., Krejtz, I. “A unified look at cultural heritage: Comparison of aggregated scanpaths over architectural artifacts”. In: *Proceedings of the ACM on human-Computer Interaction* 7.ETRA (2023), pp. 1–17.

[70] Pan, Y., Zhang, Q., Zhang, Y., Ge, X., Gao, X., Yang, S., Xu, J. “Lane-change intention prediction using eye-tracking technology: A systematic review”. In: *Applied ergonomics* 103 (2022), p. 103775.

[71] Chi, C.-F., Lin, F.-T. “A new method for describing search patterns and quantifying visual load using eye movement data”. In: *International Journal of Industrial Ergonomics* 19.3 (1997), pp. 249–257.

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

[73] Rayner, K. "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3 (1998), p. 372.

[74] Pannasch, S., Helmert, J. R., Roth, K., Herbold, A.-K., Walter, H. "Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions". In: *Journal of Eye Movement Research* 2.2 (2008).

[75] Lyu, M., Li, F., Lee, C.-H., Chen, C.-H. "VALIO: Visual attention-based linear temporal logic method for explainable out-of-the-loop identification". In: *Knowledge-Based Systems* (2024), p. 112086.

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

[77] Li, R., Hu, M., Cui, J., Wang, L., Sourina, O. "Ensemble of randomized neural network and boosted trees for eye-tracking-based driver situation awareness recognition and interpretation". In: *International Conference on Neural Information Processing*. Springer. 2023, pp. 500–511.

[78] Shi, Y., Zhu, Y., Mehta, R. K., Du, J. "A neurophysiological approach to assess training outcome under stress: A virtual reality experiment of industrial shutdown maintenance using Functional Near-Infrared Spectroscopy (fNIRS)". In: *Advanced Engineering Informatics* 46 (2020), p. 101153.

[79] Natras, R., Soja, B., Schmidt, M. "Ensemble machine learning of Random Forest, AdaBoost and XGBoost for vertical total electron content forecasting". In: *Remote Sensing* 14.15 (2022), p. 3547.

[80] Zhang, Y., Li, Q., Nahata, S., Jamal, T., Cheng, S.-K., Cauwenberghs, G., Jung, T.-P. "Integrating Large Language Model, EEG, and Eye-Tracking for Word-Level Neural State Classification in Reading Comprehension". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2024).

[81] Zhang, Y., Yang, S., Cauwenberghs, G., Jung, T.-P. "From word embedding to reading embedding using large language model, eeg and eye-tracking". In: *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2024, pp. 1–4.

[82] Cheng, B., Fan, C., Fu, H., Huang, J., Chen, H., Luo, X. "Measuring and computing cognitive statuses of construction workers based on electroencephalogram: a critical review". In: *IEEE Transactions on Computational Social Systems* 9.6 (2022), pp. 1644–1659.

[83] Qin, H., Zhu, H., Jin, X., Song, Q., El-Yacoubi, M. A., Gao, X. "EmMixformer: Mix transformer for eye movement recognition". In: *arXiv preprint arXiv:2401.04956* (2024).

[84] Suhari, A. R., Hartanto, R., Wibirama, S. "Improving performance of eye movements classification using cnn-transformer model". In: *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*. IEEE. 2023, pp. 751–756.

[85] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[86] He, H., Bai, Y., Garcia, E. A., Li, S. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Ieee. 2008, pp. 1322–1328.

[87] Brandt, J., Lanzén, E. "A comparative review of SMOTE and ADASYN in imbalanced data classification". In: (2021).

[88] Liu, T.-Y. "Easyensemble and feature selection for imbalance data sets". In: *2009 international joint conference on bioinformatics, systems biology and intelligent computing*. IEEE. 2009, pp. 517–520.

[89] Elkan, C. "The foundations of cost-sensitive learning". In: *International joint conference on artificial intelligence*. Vol. 17. 1. Lawrence Erlbaum Associates Ltd. 2001, pp. 973–978.

[90] Bian, K., Priyadarshi, R. "Machine learning optimization techniques: a survey, classification, challenges, and future research issues". In: *Archives of Computational Methods in Engineering* 31.7 (2024), pp. 4209–4233.

[91] Falkner, S., Klein, A., Hutter, F. "BOHB: Robust and efficient hyperparameter optimization at scale". In: *International conference on machine learning*. PMLR. 2018, pp. 1437–1446.

[92] Sadoune, H., Rihani, R., Marra, F. S. "DNN model development of biogas production from an anaerobic wastewater treatment plant using

Bayesian hyperparameter optimization". In: *Chemical Engineering Journal* 471 (2023), p. 144671.

[93] Platt, J. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74.

[94] Fawcett, T. "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8 (2006), pp. 861–874.

[95] Kim, Y., Wu, J., Abdulle, Y., Gao, Y., Wu, H. "Enhancing human-computer interaction in chest x-ray analysis using vision and language model with eye gaze patterns". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 184–194.

[96] Zhang, Y., Li, F., Chang, D. "VR rehabilitation system evaluator: A fNIRS-based and LLM-enabled evaluation paradigm for Mild Cognitive Impairment". In: *Advanced Engineering Informatics* 62 (2024), p. 102734.

[97] Xu, M., Niyato, D., Chen, J., Zhang, H., Kang, J., Xiong, Z., Mao, S., Han, Z. "Generative AI-empowered simulation for autonomous driving in vehicular mixed reality metaverses". In: *IEEE Journal of Selected Topics in Signal Processing* 17.5 (2023), pp. 1064–1079.

[98] Liu, Q., Li, F., Ng, K. K., Han, J., Feng, S. "Accident investigation via LLMs reasoning: HFACS-guided Chain-of-Thoughts enhance general aviation safety". In: *Expert Systems with Applications* 269 (2025), p. 126422.

[99] Li, Y., Zhang, Y., Sun, L. "Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents". In: *arXiv preprint arXiv:2310.06500* (2023).

[100] Tsai, W.-E., Liu, Y.-C. "Aisen-Web-Based Gaze-Tracking Assistive Communication Interface with Word Cards Generated by LLMs". In: *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 2023, pp. 1–3.

[101] Li, C., Zhang, D., Wang, J. "LLM-assisted Labeling Function Generation for Semantic Type Detection". In: *arXiv preprint arXiv:2408.16173* (2024).

[102] Xie, J., Li, X., Yuan, Y., Guan, Y., Jiang, J., Guo, X., Peng, X. "Knowledge-based dynamic prompt learning for multi-label disease diagnosis". In: *Knowledge-Based Systems* 286 (2024), p. 111395.

[103] Kim, C. Y., Lee, C. P., Mutlu, B. "Understanding large-language model (llm)-powered human-robot interaction". In: *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*. 2024, pp. 371–380.

[104] Wu, X., Li, Z. "A review of alarm system design for advanced control rooms of nuclear power plants". In: *International Journal of Human–Computer Interaction* 34.6 (2018), pp. 477–490.

[105] Eißfeldt, H., Grasshoff, D., Hasse, C., Hörmann, H.-J., Schulze Kissing, D., Stern, C., Wenzel, J., Zierke, O. "Aviator 2030-Ability Requirements in Future ATM Systems II: Simulations and Experiments". In: (2010).

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

[107] Liu, Y., Trapsilawati, F., Lan, Z., Sourina, O., Johan, H., Li, F., Chen, C.-H., Mueller-Wittig, W. “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*. Springer. 2020, pp. 715–726.

[108] Saward, J. R., Stanton, N. A. “Individual latent error detection: Simply stop, look and listen”. In: *Safety science* 101 (2018), pp. 305–312.

[109] Kontogiannis, T., Malakis, S. “A proactive approach to human error detection and identification in aviation and air traffic control”. In: *Safety Science* 47.5 (2009), pp. 693–706.

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

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

[112] Bruder, C., Hasse, C. “Differences between experts and novices in the monitoring of automated systems”. In: *International Journal of Industrial Ergonomics* 72 (2019), pp. 1–11.

[113] Li, F., Chen, C.-H., Lee, C.-H., Feng, S. “Artificial intelligence-enabled non-intrusive vigilance assessment approach to reducing traffic controller’s human errors”. In: *Knowledge-Based Systems* 239 (2022), p. 108047.

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

[115] Orquin, J. L., Loose, S. M. “Attention and choice: A review on eye movements in decision making”. In: *Acta psychologica* 144.1 (2013), pp. 190–206.

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

[117] Hollnagel, E. “Looking for errors of omission and commission or The Hunting of the Snark revisited”. In: *Reliability Engineering & System Safety* 68.2 (2000), pp. 135–145.

[118] Xu, J., Park, S. H., Zhang, X., Hu, J. “The improvement of road driving safety guided by visual inattentional blindness”. In: *IEEE transactions on intelligent transportation systems* 23.6 (2021), pp. 4972–4981.

[119] Sarter, N. B., Woods, D. D. “Situation awareness: A critical but ill-defined phenomenon”. In: *Situational awareness* (2017), pp. 445–458.

[120] Yu, X., Chen, C.-H., Yang, H. “Air traffic controllers’ mental fatigue recognition: A multi-sensor information fusion-based deep learning approach”. In: *Advanced Engineering Informatics* 57 (2023), p. 102123.

[121] Ferris, J. R., Tomlinson, M. A., Ward, T. N., Pepin, M. E., Malek, M. H. “Reduced electromyographic fatigue threshold after performing a cognitive fatiguing task”. In: *The Journal of Strength & Conditioning Research* 35.1 (2021), pp. 267–274.

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

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

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

[125] Diamond, J. *The pupil. Anatomy, physiology and clinical applications.* 2001.

[126] Mathôt, S. “Pupillometry: Psychology, physiology, and function”. In: *Journal of cognition* 1.1 (2018), p. 16.

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

[128] Richards, A., Hannon, E. M., Vitkovitch, M. “Distracted by distractors: Eye movements in a dynamic inattentional blindness task”. In: *Consciousness and cognition* 21.1 (2012), pp. 170–176.

[129] **Li, Z.**, Li, Z., Li, F. “Visual Attention Analytics for Individual Perception Differences and Task Load-Induced Inattentional Blindness”. In: *International Conference on Human-Computer Interaction*. Springer. 2023, pp. 71–83.

[130] Bodala, I. P., Li, J., Thakor, N. V., Al-Nashash, H. “EEG and eye tracking demonstrate vigilance enhancement with challenge integration”. In: *Frontiers in human neuroscience* 10 (2016), p. 273.

[131] Fan, J., Smith, A. P. “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*. Springer. 2017, pp. 90–105.

[132] Wilkinson, R. T., Seales, D. M. “EEG event-related potentials and signal detection”. In: *Biological psychology* 7.1-2 (1978), pp. 13–28.

[133] Navarro-Cebrian, A., Knight, R. T., Kayser, A. S. “Error-monitoring and post-error compensations: dissociation between perceptual failures and motor errors with and without awareness”. In: *Journal of Neuroscience* 33.30 (2013), pp. 12375–12383.

[134] Xu, J., Guo, K., Zhang, X., Sun, P. Z. “Left gaze bias between LHT and RHT: a recommendation strategy to mitigate human errors in left-and right-hand driving”. In: *IEEE Transactions on Intelligent Vehicles* (2023).

[135] Tse, C.-Y., Shum, Y.-H., Xiao, X.-Z., Wang, Y. “Fronto-occipital mismatch responses in pre-attentive detection of visual changes: Implication on a generic brain network underlying Mismatch Negativity (MMN)”. In: *Neuroimage* 244 (2021), p. 118633.

[136] Wolfe, J. M., Kosovicheva, A., Wolfe, B. “Normal blindness: When we look but fail to see”. In: *Trends in Cognitive Sciences* 26.9 (2022), pp. 809–819.

[137] Lleras, A., Buetti, S., Xu, Z. J. “Incorporating the properties of peripheral vision into theories of visual search”. In: *Nature Reviews Psychology* 1.10 (2022), pp. 590–604.

[138] Lee, J. H., Ostwald, M. J. “The ‘visual attractiveness’ of architectural facades: measuring visual complexity and attractive strength in architecture”. In: *Architectural science review* 66.1 (2023), pp. 42–52.

[139] Qin, L., Shi, Y., He, Y., Zhang, J., Zhang, X., Li, Y., Deng, T., Yan, H. “ID-YOLO: Real-time salient object detection based on the driver’s fixation region”. In: *IEEE Transactions on Intelligent Transportation Systems* 23.9 (2022), pp. 15898–15908.

[140] Onkhar, V., Kumaaravelu, L., Dodou, D., De Winter, J. *Towards context-aware road user safety systems: Design explorations using eye-tracking, object detection, and GPT-4V*. 2024.

[141] Hasan, M. Z., Chen, J., Wang, J., Rahman, M. S., Joshi, A., Velipasalar, S., Hegde, C., Sharma, A., Sarkar, S. "Vision-language models can identify distracted driver behavior from naturalistic videos". In: *IEEE Transactions on Intelligent Transportation Systems* (2024).

[142] Wang, X., Kim, H., Rahman, S., Mitra, K., Miao, Z. "Human-LLM collaborative annotation through effective verification of LLM labels". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 2024, pp. 1–21.

[143] Li, F., Feng, S., Yan, Y., Lee, C.-H., Ong, Y. S. "Virtual Co-Pilot: Multimodal Large Language Model-enabled Quick-access Procedures for Single Pilot Operations". In: *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE. 2024, pp. 1501–1506.

[144] Shahriar, S., Lund, B. D., Mannuru, N. R., Arshad, M. A., Hayawi, K., Bevara, R. V. K., Mannuru, A., Batool, L. "Putting gpt-4o to the sword: A comprehensive evaluation of language, vision, speech, and multimodal proficiency". In: *Applied Sciences* 14.17 (2024), p. 7782.

[145] Ma, C., Huang, J.-B., Yang, X., Yang, M.-H. "Robust visual tracking via hierarchical convolutional features". In: *IEEE transactions on pattern analysis and machine intelligence* 41.11 (2018), pp. 2709–2723.

[146] Yu, H., Kwon, M. "Altered eye movements during reading with simulated central and peripheral visual field defects". In: *Investigative Ophthalmology & Visual Science* 64.13 (2023), pp. 21–21.

[147] Ma, X., Liu, Y., Clariana, R., Gu, C., Li, P. "From eye movements to scanpath networks: A method for studying individual differences in expos-

itory text reading”. In: *Behavior research methods* 55.2 (2023), pp. 730–750.

[148] Ning, K., Liu, Y., Su, Y., Jiang, K. “Point-voxel and bird-eye-view representation aggregation network for single stage 3D object detection”. In: *IEEE Transactions on Intelligent Transportation Systems* 24.3 (2022), pp. 3223–3235.

[149] Li, Q., Ng, K. K., Simon, C., Yiu, C. Y., Lyu, M. “Recognising situation awareness associated with different workloads using EEG and eye-tracking features in air traffic control tasks”. In: *Knowledge-Based Systems* 260 (2023), p. 110179.

[150] Zhang, Y., Liao, Y., Zhang, Y., Huang, L. “Emergency braking intention detect system based on K-order propagation number algorithm: a network perspective”. In: *Brain Sciences* 11.11 (2021), p. 1424.

[151] Helton, W. S., Matthews, G., Warm, J. S. “Stress state mediation between environmental variables and performance: The case of noise and vigilance”. In: *Acta psychologica* 130.3 (2009), pp. 204–213.

[152] Tippey, K. G., Roady, T., Rodriguez-Paras, C., Brown, L. J., Rantz, W. G., Ferris, T. K. “General aviation weather alerting: the effectiveness of different visual and tactile display characteristics in supporting weather-related decision making”. In: *The International Journal of Aerospace Psychology* 27.3-4 (2017), pp. 121–136.

[153] Trapsilawati, F., Qu, X., Wickens, C. D., Chen, C.-H. “Human factors assessment of conflict resolution aid reliability and time pressure in future air traffic control”. In: *Ergonomics* 58.6 (2015), pp. 897–908.

[154] Farley, T. C., Hansman, R. J., Amonlirdviman, K., Endsley, M. R. “Shared information between pilots and controllers in tactical air traffic control”. In: *Journal of Guidance, Control, and Dynamics* 23.5 (2000), pp. 826–836.

[155] Allendoerfer, K. R., Friedman-Berg, F., Pai, S. *Human factors analysis of safety alerts in air traffic control*. Tech. rep. United States. Department of Transportation. Federal Aviation Administration …, 2007.

[156] Jürgensohn, T., Park, I., Sheridan, T. B., Meyer, J. “Conflict warnings and the search and detection of collisions in a simulated ATC-like task”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 45. 4. SAGE Publications Sage CA: Los Angeles, CA. 2001, pp. 458–462.

[157] Yang, S., Wilson, K., Roady, T., Kuo, J., Lenné, M. G. “Beyond gaze fixation: Modeling peripheral vision in relation to speed, Tesla Autopilot, cognitive load, and age in highway driving”. In: *Accident Analysis & Prevention* 171 (2022), p. 106670. ISSN: 0001-4575.

[158] Binatha, C. “Hue saturation value (HSV) color space for content based image retrieval”. In: *International Journal of Engineering Research* 5.04 (2016).

[159] Yong, G., Jeon, K., Gil, D., Lee, G. “Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model”. In: *Computer-Aided Civil and Infrastructure Engineering* 38.11 (2023), pp. 1536–1554.

[160] Chamberlain, L. “Eye tracking methodology; theory and practice”. In: *Qualitative Market Research: An International Journal* 10.2 (2007), pp. 217–220.

[161] **Li, Z.**, Li, F., Lyu, M. “Tracking the Unseen and Unaware: Deciphering Controllers’ Detection Failures to Warnings Through Eye-Tracking Metrics”. In: *International Journal of Human–Computer Interaction* (2025), pp. 1–20.

[162] Li, F., Lee, C.-H., Chen, C.-H., Khoo, L. P. “Hybrid data-driven vigilance model in traffic control center using eye-tracking data and context data”. In: *Advanced Engineering Informatics* 42 (2019), p. 100940.

[163] Harwood, M. R. *The Fourier analysis of saccadic eye movements*. University of London, University College London (United Kingdom), 2003.

[164] Yoo, Y.-J. “Fault detection of induction motor using fast Fourier transform with feature selection via principal component analysis”. In: *International Journal of Precision Engineering and Manufacturing* 20.9 (2019), pp. 1543–1552.

[165] Subramanian, H., Ibey, B. L., Xu, W., Wilson, M. A., Ericson, M. N., Coté, G. L. “An autocorrelation-based time domain analysis technique for monitoring perfusion and oxygenation in transplanted organs”. In: *IEEE transactions on biomedical engineering* 52.7 (2005), pp. 1355–1358.

[166] Silagyi II, D. V., Liu, D. “Prediction of severity of aviation landing accidents using support vector machine models”. In: *Accident Analysis & Prevention* 187 (2023), p. 107043.

[167] Zhou, P., Chen, L., Dai, X., Li, B., Chai, T. "Intelligent prediction of train delay changes and propagation using RVFLNs with improved transfer learning and ensemble learning". In: *IEEE Transactions on Intelligent Transportation Systems* 22.12 (2020), pp. 7432–7444.

[168] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. "Segment anything". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4015–4026.

[169] He, S., Bao, R., Li, J., Grant, P. E., Ou, Y. "Accuracy of segment-anything model (sam) in medical image segmentation tasks". In: *CoRR* (2023).

[170] O'Shea, D. M., Fieo, R. A. "Individual differences in fluid intelligence predicts inattentional blindness in a sample of older adults: A preliminary study". In: *Psychological research* 79 (2015), pp. 570–578.

[171] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

[172] Zhu, Q., Hou, B., Mathai, T. S., Mukherjee, P., Jin, Q., Chen, X., Wang, Z., Cheng, R., Summers, R. M., Lu, Z. "How Well Do Multi-modal LLMs Interpret CT Scans? An Auto-Evaluation Framework for Analyses". In: *arXiv preprint arXiv:2403.05680* (2024).

[173] Grippenkoven, J., Dietsch, S. "Gaze direction and driving behavior of drivers at level crossings". In: *Journal of Transportation Safety & Security* 8.sup1 (2016), pp. 4–18.

[174] Zhang, H., Xu, M. "Modeling temporal information using discrete fourier transform for recognizing emotions in user-generated videos". In: *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2016, pp. 629–633.

[175] Harezlak, K., Kasprowski, P. "Searching for chaos evidence in eye movement signals". In: *Entropy* 20.1 (2018), p. 32.

[176] Clarke, D. D., Ward, P., Bartle, C., Truman, W. "The role of motorcyclist and other driver behaviour in two types of serious accident in the UK". In: *Accident Analysis & Prevention* 39.5 (2007), pp. 974–981.

[177] **Li, Z.**, Li, F., Xu, G., Li, D. "Beyond the Gaze: Peripheral Vision-Aware Visual Detection Failures Recognition Through LLM-Based Fixation Coordinate-Sensitive Analysis". In: *IEEE Transactions on Intelligent Transportation Systems* 99 (2025), pp. 1–20.

[178] Jiang, Y., Guo, Z., Rezazadegan Tavakoli, H., Leiva, L. A., Oulasvirta, A. "EyeFormer: predicting personalized scanpaths with transformer-guided reinforcement learning". In: *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 2024, pp. 1–15.

[179] Verschooren, S., Schindler, S., De Raedt, R., Pourtois, G. "Switching attention from internal to external information processing: A review of the literature and empirical support of the resource sharing account". In: *Psychonomic bulletin & review* 26.2 (2019), pp. 468–490.

[180] Yang, L., Shami, A. "On hyperparameter optimization of machine learning algorithms: Theory and practice". In: *Neurocomputing* 415 (2020), pp. 295–316.

[181] Suguna, R., Suriya Prakash, J., Aditya Pai, H., Mahesh, T., Vinoth Kumar, V., Yimer, T. E. “Mitigating class imbalance in churn prediction with ensemble methods and SMOTE”. In: *Scientific Reports* 15.1 (2025), p. 16256.

[182] Burez, J., Van den Poel, D. “Handling class imbalance in customer churn prediction”. In: *Expert Systems with Applications* 36.3 (2009), pp. 4626–4636.

[183] Davis, J., Goadrich, M. “The relationship between Precision-Recall and ROC curves”. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 233–240.

[184] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., Hutter, F. “Efficient and robust automated machine learning”. In: *Advances in neural information processing systems* 28 (2015).

[185] **Li, Z.**, Li, F. “Vigilant Air Traffic Control: Gaze-based Recognition of Detection Failures to Visual Warnings”. In: *Advances in Human Factors of Transportation* 148.148 (2024).

[186] Zhang, Z., Li, D., Sun, Y., Lee, C.-h., Feng, S., Li, F. “How Do We Team Up? Human-Machine Co-driving Style Assessment Through Visual Dynamic Analysis and Vision-Language Model”. In: *International Conference on Human-Computer Interaction*. Springer. 2025, pp. 287–304.

[187] Liu, J., Ji, L., Chen, K., Li, C., Duan, H. “Railway operational hazard prediction and control based on knowledge graph embedding and topological analysis”. In: *Reliability Engineering & System Safety* 258 (2025), p. 110917.

Appendices

Appendix I. Warning types and examples used in DF induction experiment

The table 6.1 provides a comprehensive summary of the warning types and their representation formats utilized in the simulation experiments discussed in Section 3.2.2.

Appendix II. Sample distribution across participants

For the experimental setup in Section 4.2, 7 participants were excluded from an initial pool of 38 participants, due to invalid data: 3 for eye-tracking calibration failure, 3 for Euroscope platform timing issues preventing alert timing confirmation, and 1 for missing platform log data. Of the 494 total samples collected, 442 (89%) were valid and 52 (11%) were invalid, distributed across the remaining 31 participants as shown in Fig. 6.1. Invalid samples resulted from technical factors: (1) network delays in the Euroscope platform, causing alerts to exceed the intended 5-second duration (e.g., 6 seconds), and (2) track loss due to head movements beyond the eye tracker's range, leading to gaps in gaze data.

Table 6.1: 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.

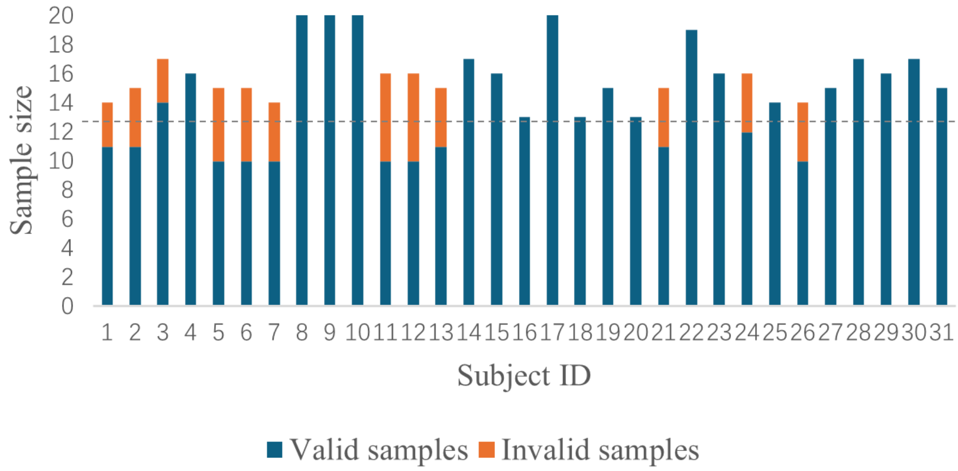


Figure 6.1: Distribution of valid and invalid samples per participant. The dashed line represents that participant received at least 13 alerts.

Appendix III. A sample output from GPT-4V Model

Output: The provided image shows a red rectangle in the upper center of the image, marked with "EMG" indicating an alert. The blue points and lines within the image represent the eye movement trajectory of an operator.

Upon close inspection:

1. Identifying red rectangle and blue trajectory:

- The red rectangle is clearly marked at the top center of the image.

- There are several blue points and lines observed in the vicinity of this red rectangle.

2. Assessing the blue points or lines inside the red rectangle:

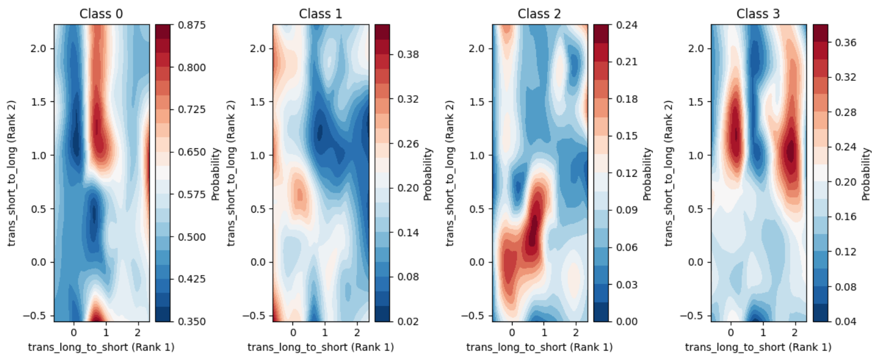
- There are clearly multiple blue points and lines (eye fixation and saccades) inside the red rectangle.

Based on the criteria provided:

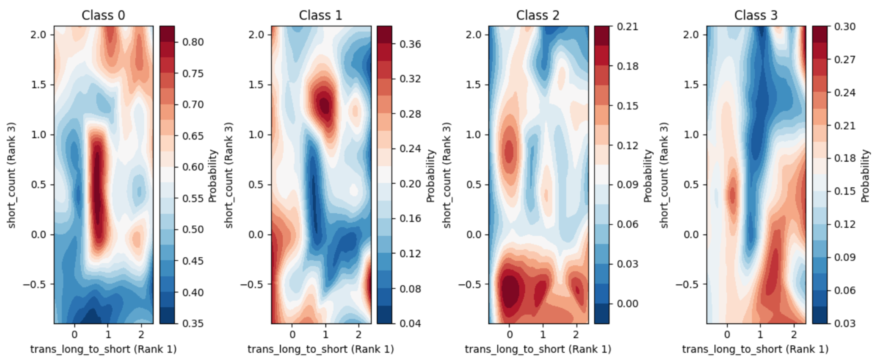
- Since there are blue points and lines inside the red rectangle, this indicates that the subject’s attention was captured by the emergency.

Final Classification: Captured alert

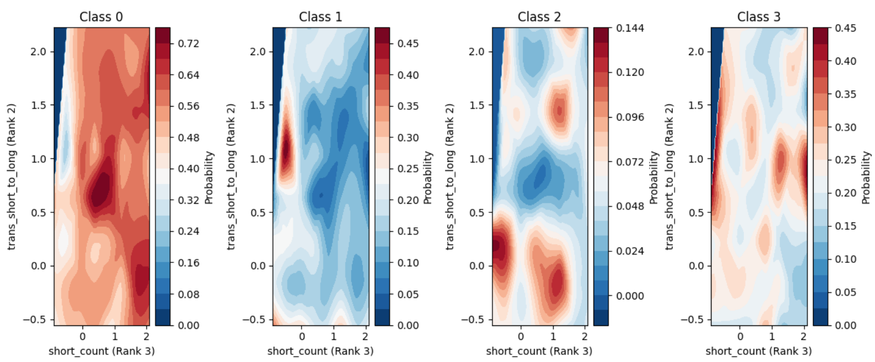
Appendix IV. 2D heatmaps of joint probabilities among top-3 features



(a) 2D heatmap of top-1 & top-2 feature pair



(b) 2D heatmap of top-1 & top-3 feature pair



(c) 2D heatmap of top-2 & top-3 feature pair

Figure 6.2: Joint probability distribution of top 1-3 feature pairs: 2D heatmap visualization in train set. Here, Class 0-3 denote Effective detection, Peripheral vision neglect, Look-but-fail-to-see error, and Pre-attentive detection, respectively.

Appendix V. Information sheet of experiment

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

INFORMATION SHEET

Unseen Horizons: A Study on Detection Failures with Air Traffic Controllers

You are invited to participate in the above project conducted by Dr Li Fan, who is a staff member of the Department of Aeronautical and Aviation Engineering in The Hong Kong Polytechnic University. The project has been approved by the Human Subjects Ethics Sub-committee (HSESC) of The Hong Kong Polytechnic University (HSESC Reference Number: HSEARS20211117002).

The aims/objectives of this project are to study the characteristics of online students' visual attention collected by Eye tracker, EEG device and capture students' attention via interaction enhancement and a reasonable online course arrangement.

You are invited to take part in a procedure to investigate the situation awareness level of the subjects. Measurements will be taken by eye tracker equipment. The eye tracker measures the eyeball positions and movement. There should be minimal discomfort or risk. The test does not hurt and the whole investigation will take about an hour.

The testing should not result in any undue discomfort, but you will need to perform the tasks required by the researchers. You are required to complete the tasks as the pilot role and perform flying activities in a flight simulator.

The information you provide as part of the project is the research data. Any research data from which you can be identified is known as personal data. Personal data does not include data where the identity has been removed (anonymous data). We will minimise our use of personal data in the study as much as possible. The researcher and his team, supervisor will have access to personal data and research data for the purposes of the study. Responsible members of The Hong Kong Polytechnic University may be given access for monitoring and/or audit of the research.

All information related to you will remain confidential. For data safety and confidentiality, the data will be stored in a locked computer. All the data, samples and study records will be stored with a password-protected zip. Only the investigator, the research staff and the research students under the investigator's supervision have the right to access the data in the period of conducting the research experiments and research project. All the data, samples and study records will be stored with a password-protected achiever. The information collected will be kept one year after project completion/publication or public release of the research results. The Hong Kong Polytechnic University takes reasonable precautions to prevent the loss, misappropriation, unauthorized access or destruction of the information you provide.

You have every right to withdraw from the study before or during the measurement without penalty of any kind.

If you have any questions, you may ask our helpers now or later, even after the study has started.

You may contact Dr Li Fan (tel. no.: 3400 2468/ email: fan-5.li@__________) of PolyU under the following situations:

a. if you have any other questions in relation to the study;
b. if, under very rare conditions, you become injured as a result of your participation in the study; or
c. if you want to get access to/or change your personal data before 30th September 2023.

In the event you have any complaints about the conduct of this research study, you may contact Miss Cherrie Mok, the Secretary of the Human Subjects Ethics Sub-Committee of The Hong Kong Polytechnic University by email (cherrie.mok@__________) or in writing (c/o Research Office of The Hong Kong Polytechnic University) stating clearly the responsible person and department of this study as well as the HSESC Reference Number.

Thank you for your interest in participating in this study.

Dr Li Fan
Principal Investigator/Chief Investigator

Hung Hom Kowloon Hong Kong 香港 九龍 紅磡
Tel 電話 (852) 2766 5111 Fax 傳真 (852) 2784 3374
Email 電郵 polyu@polyu.edu.hk
Website 網址 www.polyu.edu.hk

參與者須知

基於視覺及認知註意力分析的空中交通管制員註意力失誤研究

我們誠邀閣下參與上述由香港理工大學航空及民航工程學系李凡博士開展研究。此研究已獲香港理工大學人類實驗對象操守小組委員會批准（人類實驗對象操守小組委員會項目參考編號: HSEARS20211117002）。

此研究的目的是通過在模擬空中交通管制員的真實工作場景中，研究操控者對警告或其他關鍵信息的反應來了解他們的視覺和認知註意力分佈，分析眼動儀和腦電圖采集的特點。

過程中，研究人員將透過眼動儀和腦電圖設備測量受試者的態勢感知程度。眼動儀用於測量眼球的位置和動態。受試者於過程中或感到輕微不適或有輕微風險，但實驗不會對受試者造成傷害。整個實驗需時約一個小時。

此實驗不會引致強烈不適，但受試者需根據研究人員的指引完成實驗。受試者將以空勤人員的身份模擬處理相應指令, 並於模擬器執飛指定任務。

實驗中收集的數據為「研究數據」。任何可以識別出受試者的數據為「個人數據」。個人數據並不包括已刪除身份識別資料的數據（ 即匿名數據）。我們將盡可能減少在研究中對個人數據的使用。只有項目負責人及其團隊可存取及瀏覽研究數據和個人數據進行研究。香港理工大學的相關人士亦會被授予監察和審核此研究項目的權限。

所有與受試者有關的資料將會被保密。為保障數據的安全性和保密性，所有數據將儲存在已鎖定的電腦設備裏。所有數據，樣本和研究記錄將額外使用密碼保存。在進行實驗和研究項目期間，只有相關研究人員可在項目負責人的監督下存取數據。我們所收集的數據將於項目完成或發布研究結果後保存一年。香港理工大學會採取合理的措施防止丟失、盜用、未經授權的存取或破壞所收集到的數據。

如您在實驗開始之前或實驗期間退出研究，您將不會受到任何形式的懲罰。

如您有任何疑問，可隨時向研究人員查詢。

若有以下情況，您可聯繫李凡博士（電話：34002468／電子郵件：fan-5.li@）：

a. 如果您有其他與此研究有關的問題；
b. 如果在非常罕見的情況下，您因參加此實驗而受傷；
c. 如果您想在 2023 年 9 月 30 日之前存取或更改您的個人數據。

如果您對此項研究有任何投訴，可透過電子郵件（cherrie.mok@　　　　　　)或郵件聯繫香港理工大學人類實驗對象操守小組委員會秘書莫小姐。若郵寄投訴，請於信封註明「轉交香港理工大學研究事務處」、研究負責人、及其所屬部門及人類實驗對象操守小組委員會項目參考編號。

我們再次感謝您的參與。

李凡博士
項目負責人

Hung Hom Kowloon Hong Kong 香港 九龍 紅磡
Tel 電話 (852) 2766 5111 Fax 傳真 (852) 2784 3374
Email 電郵 polyu@polyu.edu.hk
Website 網址 www.polyu.edu.hk

Appendix VI. Consent form of experiment

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

CONSENT TO PARTICIPATE IN RESEARCH

Unseen Horizons: A Study on Detection Failures with Air Traffic Controllers

I ________________________ hereby consent to participate in the captioned research conducted by Dr Li Fan.

I understand that information obtained from this research may be used in future research and published. However, my right to privacy will be retained, i.e. my personal details will not be revealed.

The procedure as set out in the attached information sheet has been fully explained. I understand the benefit and risks involved. My participation in the project is voluntary.

I acknowledge that I have the right to question any part of the procedure and can withdraw at any time without penalty of any kind.

Name of participant ____

Signature of participant ____

Name of Parent or
Guardian (if applicable) ____

Signature of Parent
or Guardian (if applicable) ____

Name of researcher ____

Signature of researcher ____

Date ____

Hung Hom Kowloon Hong Kong 香港 九龍 紅磡
Tel 電話 (852) 2766 5111 Fax 傳真 (852) 2784 3374
Email 電郵 polyu@polyu.edu.hk
Website 網址 www.polyu.edu.hk

THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

參與研究同意書

基於視覺及認知注意力分析的空中交通管制員註意力失誤研究

本人___________________同意參與由李凡博士開展的上述研究。

本人知悉此研究所得的資料可能被用作日後的研究及發表，但本人的私隱權利將得以保留，即本人的個人資料不會被公開。

研究人員已向本人清楚解釋列在所附資料卡上的研究程序，本人明瞭當中涉及的利益及風險；本人自願參與研究項目。

本人知悉本人有權就程序的任何部分提出疑問，並有權隨時退出而不受任何懲處。

參與者姓名 ______________________________________
參與者簽署 ______________________________________
家長或監護人(如適用) 姓名 _________________________
家長或監護人(如適用) 簽署 _________________________
研究人員姓名 ____________________________________
研究人員簽署 ____________________________________
日期 ____

Hung Hom Kowloon Hong Kong 香港 九龍 紅磡
Tel 電話 (852) 2766 5111 Fax 傳真 (852) 2784 3374
Email 電郵 polyu@polyu.edu.hk
Website 網址 www.polyu.edu.hk