Advanced Engineering Informatics 55 (2023) 101894

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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

Full length article

# Predicting separation errors of air traffic controllers through integrated sequence analysis of multimodal behaviour indicators

Ruoxin Xiong [a], Yanyu Wang [a], Pingbo Tang [a,*], Nancy J. Cooke [b], Sarah V. Ligda [b], Christopher S. Lieber [b], Yongming Liu [c]

[a] Department of Civil and Environmental Engineering, Carnegie Mellon University, 15213 PA, USA
[b] Human Systems Engineering, The Polytechnic School, Arizona State University, 85212 AZ, USA
[c] School for Engineering of Matter, Transport, and Energy, Arizona State University, 85287 AZ, USA

ARTICLE INFO

*Keywords:*
Air traffic control
Behavior indicators
Separation errors
Multi-step prediction
Multimodal data
Loss of Separation (LoS)

ABSTRACT

Predicting separation errors in the daily tasks of air traffic controllers (ATCOs) is essential for the timely implementation of mitigation strategies before performance declines and the prevention of loss of separation and aircraft collisions. However, three challenges impede accurate separation errors forecasting: 1) compounding relationships between many human factors and control processes require sufficient operation process data to capture how separation errors occur and propagate within controller-in-the-loop processes; 2) previous human factor measurement approaches are disruptive to controllers' daily operations because they use invasive sensors, such as electroencephalography (EEG) and electrocardiography (ECG), 3) errors accumulated in using the tasks and human behaviors for estimating system dynamics challenge accurate separation error predictions with sufficient leading time for proactive control actions. This study proposed a separation error prediction framework with a long leading time (>50 s) to address the above challenges, including 1) a multi-factorial model that characterizes the inter-relationships between task complexity, behavioral activity, cognitive load, and operational performance; 2) a multimodal data analytics approach to non-intrusively extract the task features (i.e., traffic density) from high-fidelity simulation systems and visual behavioral features (i.e., head pose, eyelid movements, and facial expressions) from ATCOs' facial videos; 3) an encoder-decoder Long Short-Term Memory (LSTM) network to predict long-time-ahead separation errors by integrating multimodal features for reducing accumulated errors. A user study with six experienced ATCOs tested the proposed framework using the Phoenix Terminal Radar Approach Control (TRACON) simulator. The authors evaluated the model performance through two types of metrics: 1) point-level metrics, including precision, recall, and F1-score, and 2) sequence-level metrics, including alignment accuracy and sequence similarity. The results showed that 1) the model using the task and visual behavioral features significantly improved the prediction performance compared to the model using one single feature (eyelid movements), with an improvement of up to 26.95% in alignment accuracy for 10s-ahead prediction; 2) the model that combined task and visual behavioral features had a higher or comparable performance to models with different hybrid features, achieving an alignment accuracy of 82.38% for 50s-ahead error prediction; and (3) the proposed method outperformed three baseline models – Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and classic LSTM – by 8.21%, 3.47%, and 3.14% in alignment accuracy, respectively, for predicting 50s-ahead separation errors. These results suggest that the proposed model can effectively predict separation errors in air traffic control.

## 1. Introduction

Air traffic controllers (ATCOs) play a curial role in organizing and coordinating safe, orderly, and efficient traffic flow for the airspace system. ATCOs work in approach control facilities, control towers, or route centers to prevent conflicts or collisions between aircraft. They must constantly handle a large amount of information, such as weather reports, radar screens, flight strips, and radio communications with

* Corresponding author.
*E-mail addresses:* ruoxinx@andrew.cmu.edu (R. Xiong), yanyuwan@andrew.cmu.edu (Y. Wang), ptang@andrew.cmu.edu (P. Tang), nancy.cooke@asu.edu (N.J. Cooke), sligda@asu.edu (S.V. Ligda), clieber1@email.arizona.edu (C.S. Lieber), Yongming.Liu@asu.edu (Y. Liu).

https://doi.org/10.1016/j.aei.2023.101894
Received 12 September 2022; Received in revised form 4 January 2023; Accepted 20 January 2023
Available online 8 February 2023

pilots. This profession requires specialized knowledge, short-term memory, and daily real-time decision-making, making it one of the most mentally challenging occupations [1]. Despite the high standards of operational performance required, ATCOs are prone to separation errors, which can lead to Loss of Separation (LoS) events - instances where controllers fail to maintain the required minimum distance between two aircraft. LoS events can have serious consequences, including injuries or fatalities, as seen in several aviation accidents and incidents such as the 2002 Uberlingen mid-air collision [2], the 2001 Japan Airlines mid-air incident [3], and the 1996 Charkhi Dadri mid-air collision [4].

On the other hand, as air traffic continues to grow, separation errors made by controllers have also increased. According to the Federal Aviation Administration (FAA), there were 4394 separation errors made by controllers who handled 132 million flights in 2012, more than twice the number of mistakes reported in 2011 (1895) and more than three times the number counted in 2009 (1234) [5]. Nearly 90% of separation errors made by controllers resulted from human factors rather than procedural or equipment deficiencies [6]. Controllers with high workloads or poor mental states (e.g., fatigue or distraction) often struggle to maintain situational awareness and may make severe separation errors. [7]. For this reason, identifying risk factors and modeling controllers' separation errors are essential for mitigating safety events and ensuring the safety and efficiency of national airspace systems.

Previous studies have identified several leading causes of controller-related separation errors, including cognitive load, complex air traffic situations, and poor mental states (e.g., distraction, stress, and fatigue) [8,9]. It is important to recognize that human factors in an air traffic control environment are rarely isolated - for example, a controller who is tired or stressed may have difficulty handling scenarios with high traffic density. Identifying and extracting the interactions between these factors is crucial for predicting separation errors. However, the quantitative impacts of multiple human and process characteristics on separation errors are not fully understood [10]. Sufficient process and human behavioral data are necessary to understand the interactions between these factors and how separation errors occur and propagate.

Recent studies have focused on using physiological methods to recognize spontaneous physiological activity and relate it to human factors such as cognitive load and mental states in air traffic control. These methods continuously measure various physiological activities, such as heart rates, brain activities, and eye movements, using sensors like electroencephalography (EEG), electrocardiography (ECG), and functional Near-Infrared (fNIR) spectroscopy sensors [11-15]. However, these physiological measurements can be disruptive to controllers' task performance. There has been growing interest in using computer vision to non-intrusively extract and assess human behaviors in real time. Changes in behavioral features such as head, eye, and facial expressions can be used to identify mental conditions such as fatigue, drowsiness, stress, and distraction [14,16,17]. This observation has led to the investigation of a broader range of potential human behavioral indicators for identifying ATCOs' separation errors.

ATCOs' separation error generation and propagation are dynamic processes in the spatial–temporal domains. While some recent studies have attempted to model controllers' separation errors using task-related information [18-20], the limited integration of relevant features hinders the accuracy and long-term forecasting of separation errors for ATCOs. Early prediction (i.e., multi-step prediction) of potential separation errors can provide enough leading time for preventive actions before system performance declines. Long-term forecasting requires an in-depth identification and extraction of the spatiotemporally correlated task and human factors.

In this study, the authors conducted controller-in-the-loop experiments to simulate ATCOs' operational processes and develop models to predict their separation errors. Separation errors occur when controllers fail to maintain the required minimum distance between two or more aircraft. The approach involves developing a multi-factorial model that characterizes the error-generation process of ATCOs by outlining the relationships between contributing factors and indicators of operational errors. Then, a multimodal system based on the multi-factorial model was used to extract air traffic situations and controllers' behaviors in simulated air traffic control tasks. These monitored visual behaviors include head pose, eyelid movements, and facial expressions. Finally, an encoder-decoder long short-term memory (LSTM) network was proposed to predict $n$-step-ahead separation errors ($n$ = 2, 5, 8, and 10, corresponding to 10 s, 25 s, 40 s, and 50 s as each sampling step was 5 s in the simulation experiments) based on the multimodal features obtained. This new framework has the potential to help ATCOs prevent collisions between aircraft by proactively generating a time-ahead alert of potential separation errors.

The main contributions of this research are:

- A multi-factorial model captures interactions between performance-influencing factors and indicators (e.g., task complexity, behavioral activity, and cognitive load) and shapes the operational error-generation process. This model allows for the quantitative modeling of the interactions between the relevant task and human factors and enables multimodal data analytics to predict potential separation errors.
- A multimodal sensing system simultaneously collects and exacts real-time task factors and human behaviors that characterize ATCOs' operational performance in a non-intrusive manner. This study aims at minimum intrusive human state measurements while maintaining the spatiotemporal resolution of human behaviors. The approach involves identifying a comprehensive list of behavioral activities associated with ATCOs' separation errors and characterizing the contributions of different behavioral indicators in helping predict controllers' operational performance. Such characterization can also provide insight into how similar tasks can use specific behavioral indicators to comprehend human errors.
- An encoder-decoder LSTM network can use multiple features to predict separation errors in air traffic control. Accurate early prediction for separation errors can provide enough time for taking preventive actions before system performance and safety deteriorate. The proposed model was tested in different task scenarios using a user study with six professional ATCOs using the Phoenix Terminal Radar Approach Control (TRACON) simulator. Combining multiple visual features through the new encoder-decoder LSTM, which can handle multivariate sequences with variable lengths, yields a more robust and accurate performance characterization than a single feature. The authors also compared the proposed methods with three baseline deep learning methods, including the Convolutional Neural Network (CNN), Gated Recurrent Unit (GRU), and classic LSTM, and found that the new encoder-decoder LSTM method was effective and robust in predicting separation errors for controllers.

The organization of this paper is as follows. Section 2 reviews relevant research studies that examine factors influencing controllers' operational performance and human factor measurements for predicting separation errors. Section 3 describes a multi-factorial model that characterizes the error-generation process for ATCOs and illustrates the process of multimodal feature extraction. It also proposes a model that can predict a sequence of separation error occurrences to allow preventive actions with several steps in advance (e.g., 5 s per operation step, 10-step-ahead). Section 4 presents the experimental design for collecting multimodal data from high-fidelity air traffic control. Section 5 then presents the factor analysis of the proposed multi-factorial model and evaluates the developed encoder-decoder LSTM prediction model under different input features. The authors also compare the model performance with three baseline deep learning methods for predicting multi-step separation errors. Section 6 discusses the practical implications of the proposed method, as well as the limitations and further improvements, followed by a summary of the research findings in

Section 7.

## 2. Literature review

This section reviewed the factors affecting ATCOs' operational performance, examined existing techniques for measuring human factors, and discussed approaches for modeling ATCOs' separation errors.

### 2.1. *Factors influencing controllers' operational performance*

The air traffic control system is human-centered, as ATCOs issue clearances and instructions to prevent or resolve conflicts based on processing an information stream. ATCO is one of the most mentally demanding jobs that consistently require maintaining a high level of operational performance. Human factors have been repeatedly found to impact ATCOs' operational performance and are primary determinants of separation errors. Consequently, identifying the key risk factors of ATCOs and gaining in-depth insights into their influences is essential in predicting the ATCOs' operation errors, ensuring the safety and efficiency of air traffic control.

Cognitive load, complex air traffic situations, and abnormal mental states (e.g., distraction and fatigue) are the leading causes of controller-related separation errors based on studies from ATCOs' interviews, aviation incident/accident reports, and air traffic control simulations [7,9,21]. Cognitive load characterizes the mental effort ATCOs take to accomplish traffic control tasks. High levels of cognitive load can negatively affect controller performance and increase the separation error probabilities. For example, an increased cognitive load may cause ATCOs to misjudge or overlook an unsafe developing situation, leading to LoS events. Another factor that impacts ATCO performance is the complexity of the air traffic management task. Task complexity, which is determined by the air traffic control task itself, is distinct from the cognitive load. Many studies have attempted to identify indicators of task complexity, such as traffic density and sector sizes. This study reviewed the task complexity indicators in mid-air traffic control and summarized a list of complexity factors, as shown in Table 1.

Task complexity indicators can be divided into three categories: traffic characteristics, airspace characteristics, and off-nominal events. Traffic characteristics, such as traffic density, traffic in climb or descent, and variance of ground speed, capture the aircraft's complexity dimensions in the airspace. Airspace characteristics, such as sector sizes and structure, tend to reflect the procedures-related complexity. The dimensions in the third category, off-nominal events (e.g., restricted flight areas, runway switches, and bad weather), could also impact the task complexity. For example, restricted flight airspace can indirectly affect traffic characteristics. Among many potential task complexity indicators, the number of aircraft under control (i.e., traffic density) has shown the most evident relationship to ATCOs' workload and operational performance in previous studies [21]. Typically, air traffic situations with higher traffic density often contain more difficult-to-detect air traffic conflicts, which can increase the likelihood of errors by ATCOs. For example, the FAA proposed a "staffs to traffic" strategy to match the number of ATCOs with traffic volume and related workload in a specified time and sector [22]. Following previous studies [23,24], this study designed different complexity levels of air traffic management (ATM) scenarios considering the traffic densities and potential impact of off-nominal conditions on human performance in the Next Generation (NextGen) Air Transportation System [25].

**Table 1**
Classification of the reviewed task complexity indicators.

| Categories | Task complexity indicators | References |
|---|---|---|
| Traffic characteristics | Traffic density, traffic in climb or descent, the variance of aircraft speed, etc. | [20,21,23,26] |
| Airspace characteristics | Sector size, sector structure, etc. | [27,28] |
| Off-nominal events | Weather, Loss of Radio Communication (NORDO), runway switch, restricted flight areas, minimum fuel reported, etc. | [25,29] |

Abnormal mental states, such as distraction and fatigue, can degrade the operational performance of ATCOs. For instance, interruptions or distractions within the control room, such as visual distractions from radio communications, equipment alarms, phone calls, and poor workstation layout, can distract controllers' attention from a potential conflict or cause them to forget to take preventive actions. Fatigue also influences a controller's working capacity and situation awareness. For example, the FAA investigated 3,268 ATCOs and reported that 56% of these ATCOs identified fatigue as a significant factor contributing to separation error occurrences [30]. Most previous research has focused on the effects of individual factors on ATCOs' separation errors. However, human factors that contribute to operational errors typically interact with the context of the control environment. For example, an ATCO experiencing fatigue or stress may not be able to effectively handle air traffic situations with high traffic density. The quantitative impacts of multi-factor interactions on performance are still unclear [10]. One potential solution is to develop a multi-factorial model that integrates risk factors to characterize ATCOs' operational performance. This model could enable the quantitative modeling of the interactions between the relevant task and human factors and facilitate the use of multimodal data analytics to predict potential separation errors.

### 2.2. *Human factor measurements for air traffic controllers*

Modeling and predicting ATCOs' separation errors requires the in-depth identification and extraction of human factors in forming and propagating separation errors. Existing methods that measure human factors for ATCOs mainly involve subjective and physiological measures. Table 2 summarizes the existing methods that measure the aforementioned human factors for ATCOs.

Subjective measures are commonly used to evaluate ATCOs' workload and mental state. Various survey instruments can quantify ATCOs' self-rated levels of cognitive load, fatigue, and stress during or immediately after completing a task. Examples of tools that use subjective measures to assess cognitive load in air traffic management tasks include the NASA Task Load Index (TLX) [31] and the Instantaneous Self-Assessment (ISA) [32]. While subjective measures are easy to use, they have some potential limitations [23], such as the possibility of unwillingness to report and subjective biases. Additionally, these subjective methods can be intrusive and may interfere with controllers' ongoing tasks.

Recent studies have focused on the physiological methods to recognize spontaneous physiological activities and relate them to human factors (e.g., cognitive load and mental states) in air traffic control. These methods use physiological sensors to measure various physiological activities, such as heart rate, blood pressure, brain activity, and eye movements. For example, Vogt et al. [11] measured the heart rate

**Table 2**
Classification of the reviewed human factor measurements for air traffic controllers.

| Influencing factors | Human factor measurements | Measuring equipment | References |
|---|---|---|---|
| Cognitive load | NASA TLX, ISA, heart rates, blood pressure, brain activities, eye movements | Self-report questionnaires, EEG, ECG, fNIR, eye tracker | [11,31-34] |
| Stress | Stress perception reports, brain activities, heart rates | | [35,36] |
| Fatigue | Fatigue questionnaire, brain activities, eye movements | | [37-39] |

and blood pressure to evaluate the controllers' mental workload in en-route and tower air traffic control. Dasari et al. [8] investigated the correlations between EEG data and fatigue in simulated air traffic control tasks. However, these physiological measurements may cause prolonged usability and comfort issues and interfere with controllers' daily operations.

Recent advances in computer vision show the potential for non-intrusively assessing the spatiotemporal patterns of human behavior in real time. Changes in behavioral features such as head pose, eye movements, and facial expressions can be used to identify certain mental conditions, such as fatigue, drowsiness, stress, and distraction [40-42]. Computer vision-based approaches can help characterize and identify abnormal human mental states from visual behaviors without the need for intrusive devices on the body. Table 3 shows a compilation of visual behavioral metrics used in various cognitive tasks. Although these studies use different schemes to measure workload, fatigue, distraction, drowsiness, or stress, typical visual behavior metrics include head pose, eye movements, and facial expressions. For example, Liu et al. [43] monitored drivers' facial expressions to indicate mental fatigue. Zhao et al. [14] proposed a driving distraction detection method based on drivers' head poses. However, current efforts mainly focus on using a single visual feature to characterize mental states, which can be ambiguous and may not always accurately indicate mental conditions if used alone. Additionally, few studies have investigated the relationships between operational performance and visual behaviors. Therefore, this study will integrate multimodal visual and non-visual features to examine whether human behavior activities can help predict the controllers' separation errors in air traffic control tasks.

### 2.3. Separation error prediction for air traffic controllers

Human factors, such as cognitive load and abnormal mental states, can indicate pending separation errors and negatively influence controllers' operational performance in air traffic control environments. Traditional studies have focused on identifying demanding scenarios or detecting controllers' abnormal mental states, such as fatigue, stress, and distraction [8,35]. However, the interactions between human factors and their quantitative impacts on controllers' separation errors remain unclear. Some recent studies [18,20] have attempted to associate controllers' separation errors with task information. However, the limited integration of relevant features hinders the accurate and long-term forecasting of separation errors. Multi-step-ahead prediction of separation errors would provide sufficient leading time for preventive actions before system performance declines.

Multiple-step prediction is more challenging than one-step forward prediction – predicting a sequence of steps having LoS vs. the next step of having LoS. The former requires in-depth identification and extraction of correlated tasks and involved human factors. Traditional machine learning models, such as Random Forest and Support Vector Machine (SVM), are insufficient for extracting the characteristics of sequences and representing the complex influence of multiple human and task factors over operational processes [58]. Modern deep learning models, such as CNN with more complex neural network architectures, could learn a high-dimensional complex function through a series of non-linear transformations from input data to output labels. Recurrent Neural Network (RNN) models or RNN variants (e.g., LSTM and GRU) can use their internal states to process input sequences, which helps to capture more temporal dependencies in sequences [40,59]. However, these methods may struggle to handle long-term predictions based on multivariate dependencies in input sequences.

The emergence of encoder-decoder architectures has significantly advanced the ability to model sequence-to-sequence problems, such as trajectory predictions and remaining useful life forecasts [60,61]. In this architecture, the encoder processes the input sequence and captures the complex dependencies between the inputs and the targets using a deep neural network (e.g., LSTM). The decoder then uses another deep neural network to generate the target output sequence from a fixed vector, allowing for the prediction of arbitrary length sequences. This architecture is more flexible and scalable than simple RNN architectures, which struggle with long-range modeling dependencies. Nevertheless, the encoder-decoder model has not yet been widely investigated and applied to human error identification problems for two main reasons: (1) a lack of deep understanding of complex human error generation and propagation mechanism; and (2) a lack of sufficient operation process data to identify possible factors that contribute to human errors. In this study, controller-in-the-loop simulations were used to identify a range of behavioral activities related to separation errors made by ATCOs. An encoder-decoder LSTM deep learning model was then developed to analyze multivariate time series data, including traffic and behavioral features, and make multi-step predictions of separation errors.

**Table 3**
Classification of the reviewed human factors measurements using visual behaviors.

| Human factors | Tasks/scenarios | Visual behavioral measures | References |
|---|---|---|---|
| Fatigue | Driving, air traffic control, equipment operation | Head pose, mouth movement, facial expression, eyelid movement | [43-46] |
| Distraction | Driving, learning analysis | Head pose, facial expression, gaze | [14,47,48] |
| Drowsiness | Driving | Head pose, eyelid movement, facial expression | [49-51] |
| Stress | Driving, learning analysis | Head pose, facial expression, eye gaze | [52-55] |
| Cognitive load | Driving, surgery | Facial expression, eyelid movement | [56,57] |

## 3. Methodology

This section presents a multimodal data analytics framework that uses multiple data sources to build a model for predicting separation errors. This framework involves the extraction and combination of the task contexts (e.g., air traffic density) and human features (e.g., head pose, eyelid movement, and facial expression), as shown in Fig. 1. In particular, Section 3.1 describes the multi-factorial model of factor interactions on operational performance in air traffic control. This section is necessary for understanding and conceptualizing how separation errors occur and propagate. Section 3.2 then examines the multimodal data sources to characterize critical factors that impact ATCOs' separation errors and details the multimodal feature extraction derived from various data sources. Finally, Section 3.3 proposes an encoder-decoder LSTM model for predicting multi-step separation errors with fused multimodal features, which provides ATCOs with a progressive sequence of LoS probabilities.

### 3.1. Multi-factorial model for air traffic controller operational performance

The developed multi-factorial model in Fig. 2 captures how interrelations between multiple factors influence the ATCOs' operational performance. The model illustrates the relationships between task complexity, cognitive load, abnormal mental states, and operational performance.

- **Task complexity** refers to the inherent difficulty of air traffic situations presented to ATCOs tasks [62]. Following previous studies [23,24], the authors defined different air traffic complexity levels by adjusting air traffic scenarios in terms of traffic density and off-nominal scenarios. The authors assumed that task complexity increases as aircraft density increases, thus demanding a higher mental

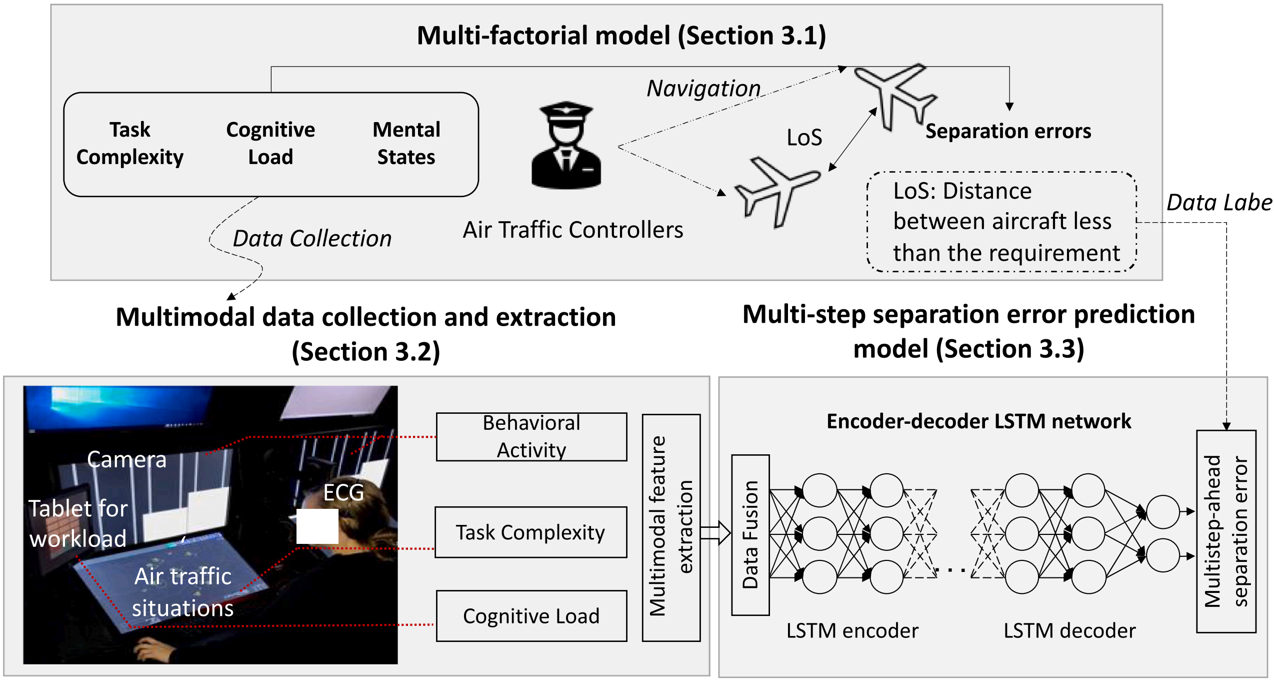

Multi-factorial model (Section 3.1)
Task Complexity
Cognitive Load
Mental States
Navigation
LoS
Separation errors
Air Traffic Controllers
Data Collection
LoS: Distance between aircraft less than the requirement
Data Labels
Multimodal data collection and extraction (Section 3.2)
Multi-step separation error prediction model (Section 3.3)
Camera
Tablet for workload
ECG
Air traffic situations
Behavioral Activity
Task Complexity
Cognitive Load
Multimodal feature extraction
Encoder-decoder LSTM network
Data Fusion
LSTM encoder
LSTM decoder
Multistep-ahead separation error

**Fig. 1.** The proposed framework of separation error prediction using multimodal behavior features.

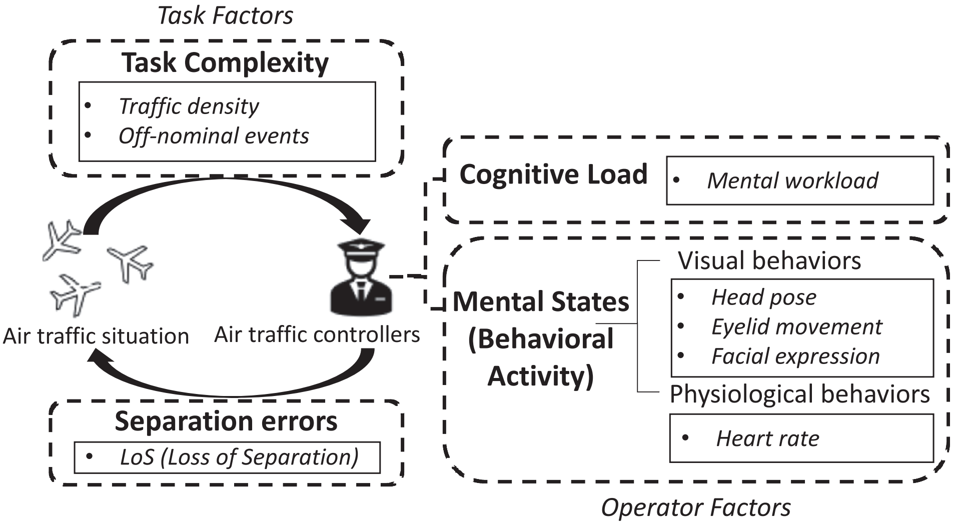


**Fig. 2.** The multi-factorial model of separation errors in air traffic control.

workload and compromising operational performance. The induced off-nominal events (e.g., runway switch and communication failures) will also increase the task complexity of ATCOs during the simulations.

- **Cognitive load** quantifies the mental effort required by ATCOs to manage air traffic situations. Task complexity is a critical driver of cognitive load [21]. In their work, controllers must pay attention to various information sources (e.g., radar screens and quartz sector maps). Higher levels of cognitive load can reduce a controller's ability to handle air traffic and increase the risk of separation errors.
- **Mental states** like fatigue, distraction, and stress, can negatively affect the performance of ACTOs in their tasks. It is believed that the controllers' visual and physiological behaviors can reveal their mental states. For instance, the head pose may indicate distraction or fatigue, while eyelid movement could reflect drowsiness and fatigue. In addition to eye and head movements, facial expressions can also convey a person's cognitive states - for example, a person's face may appear expressionless for a prolonged period when they are experiencing fatigue. Additionally, physiological behaviors like heart rate variability may be used to predict fatigue and stress and to measure mental workload [63].
- **Operational performance** assesses a controller's ability to effectively organize and coordinate safe and efficient traffic flow in dynamic airspace systems. Previous studies [20,64] have used separation errors, which occur when controllers fail to maintain the minimum separation distance of two or more aircraft, as a metric to evaluate ATCOs' operational performance. The separation error is a widely accepted method for measuring operational performance.

This model investigates the relationships between human factors (i.e., cognitive load and abnormal mental states), task factors (i.e., task complexity), and separation errors. In particular, task complexity and cognitive load can interact and change dynamically as controllers encounter different tasks in various scenarios. These changes in task complexity may lead to changes in cognitive load and contribute to separation errors. Abnormal mental states, like fatigue, stress, and distraction, can also negatively impact the operational performance of ATCOs. Considering that human activities, such as visual and physiological behaviors, can reveal these mental states, this study examines a comprehensive list of visual behavioral activities associated with operational performance. The multi-factorial model allows for the quantitative modeling of the interactions between the relevant task and human factors, which can then be used in multimodal data analytics to predict potential ATCOs' operational errors.

### 3.2. *Multimodal data collection and extraction*

The proposed multi-factorial model for ATCOs' operational performance aims to guide data collection and feature extraction to study the interactions between different factors and to build quantitative models that can predict separation errors using real-time task factors and human information. Collecting separation errors from an actual air traffic control environment is not practical and scalable due to the safety–critical nature of these events and the challenging conditions (e.g., poor mental states and complex air traffic situations) in which they occur. Instead, the proposed approach involves collecting multimodal data in different task scenarios through high-fidelity controller-in-the-loop simulation experiments. Table 4 summarizes the multimodal data sources and features that are used to characterize separation errors in air traffic control. These data include simulator tracklogs, which record airspace information (e.g., flight ID, aircraft type, and aircraft states) during the simulated air traffic control, as well as self-report ratings of mental workload and heart rate data collected using ECG equipment. The authors also used computer vision algorithms to extract visual behaviors (e.g., head poses, eyelid movements, and facial expressions) from video data. The following sections provide more technical details on these data sources and features.

#### 3.2.1. *Features of air traffic control tasks*

For this study, the authors defined different task complexity levels by adjusting air traffic scenarios in terms of traffic density and off-nominal events. Fig. 3 illustrates the traffic conditions in different simulation scenarios: (1) baseline scenarios have lower traffic density (around 2–5 aircraft at a time), (2) nominal scenarios have higher traffic density (around 6–12 aircraft at a time), and (3) off-nominal scenarios have a similar traffic density to nominal scenarios, but also introduce off-nominal events, such as turbulence reports, NORDO, active runway changes, and aircraft with minimum fuel reports.

While more task features would be desirable, our previous research of aircraft characteristics (i.e., traffic density and the number of turning aircraft) [20] has shown that traffic density is the primary indicator of ATCOs' separation errors. In contrast, the number of turning aircraft contributes little to predicting operational performance. Therefore, this study will focus on analyzing the effects of traffic density and off-nominal events.

**Table 4**
Summary of involved elements, quantitative indicators, and measuring data sources.

| Involved elements | Quantitative indicators | | Measuring data sources |
|---|---|---|---|
| Task complexity | Traffic density and distraction events | | Simulator tracklogs |
| Cognitive load | Mental workload | | Self-report ratings |
| Behavioral activity | Visual behaviors | Head poses<br>Eyelid movementFacial expressions | Video data |
| | Physiological behaviors | Heart rates | ECG data |
| Separation errors | LoS | | Simulator tracklogs |

#### 3.2.2. *Visual behavioral measures*

Visual human behaviors, such as head pose, eyelid movement, and facial expression, could reflect mental states and help to predict separation errors in air traffic control. To capture these behaviors, the authors used computer vision algorithms to extract data from surveillance videos.

(1) Head pose estimation

This study used OpenFace, an open-source toolbox developed by Baltrusaitis et al. [65], to estimate the three-dimensional (3D) head orientation and position of controllers. The algorithm first localizes the facial region using the Histogram of Oriented Gradient (HOG) features and the SVM. The Conditional Local Neural Fields (CLNF) model [66] then identifies facial landmarks with two components: (1) a point distribution model that represents landmark geometric variations, and (2) a probabilistic patch expert that describes the local shape variations of each landmark. This facial landmark detector can generate a set of 2D facial features such as the jaw outline, nose tip, eye corners, and mouth corners.

The next step is to retrieve the 3D head pose based on the 2D-3D correspondence. As in the original study [65], the 3D coordinates of six facial landmarks (nose tip, chin, left and right corners of the lip, the left corner of the left eye, and the right corner of the right eye) serve as reference landmarks to facilitate the matching process. The 3D head pose can be characterized by three degrees of freedom: pitch, roll, and yaw angles, which describe the orientation and transformation of a person's head relative to the camera. The details of 2D-3D head pose transformation can be found in the original study [65]. Fig. 4 shows head pose estimation results from a participant in the off-nominal scenario. The algorithm can measure the real-time 3D head orientation and position of air traffic controllers.

(2) Eyelid movement detection

Another algorithm calculates the widely used eye aspect ratio (EAR) [67], which is the ratio of the height to the width of the eye region, to estimate the state of eye-opening. The model uses the average EARs of the left and right eye in each video frame. When the eyes close, the pupils become occluded by the eyelids, resulting in an EAR value close to zero. Eq. (1) shows the EAR metrics based on six eye landmarks. Fig. 5 below illustrates the locations of the six facial landmarks in both open and closed eye states.

$$EAR = \frac{||p_2 - p_6|| + ||p_3 - p_5||}{2||p_1 - p_4||} \tag{1}$$

where $p_1$ and $p_4$ measure the width of the eyes while $p_2$, $p_3$, $p_5$, and $p_6$ measure the height of the eyes.

(3) Facial expression recognition

A CNN-based algorithm developed by Arriaga et al. [68] can perform real-time facial expression recognition and classification. This facial recognition model can classify faces into seven emotional categories: anger, disgust, fear, happy, sadness, surprise, and neutral. For the purpose of this study, these seven facial expression categories were further grouped into three classes: positive (happy), negative (angry, disgust, fear, sad, surprise), and neutral (neutral) to simplify the feature representation and fusion.

Fig. 6 shows the architecture of the implemented algorithm, consisting of four stacked residual depth-wise separable convolutions. The residual depth-wise separable convolution combines residual modules and depth-wise separable convolutions. Residual modules address the degradation problem by using skip connections and deeper network layers [69], while depth-wise separable convolutions reduce the number of network parameters while improving representation efficiency within a convolutional layer [70]. The last layer of this model uses a global

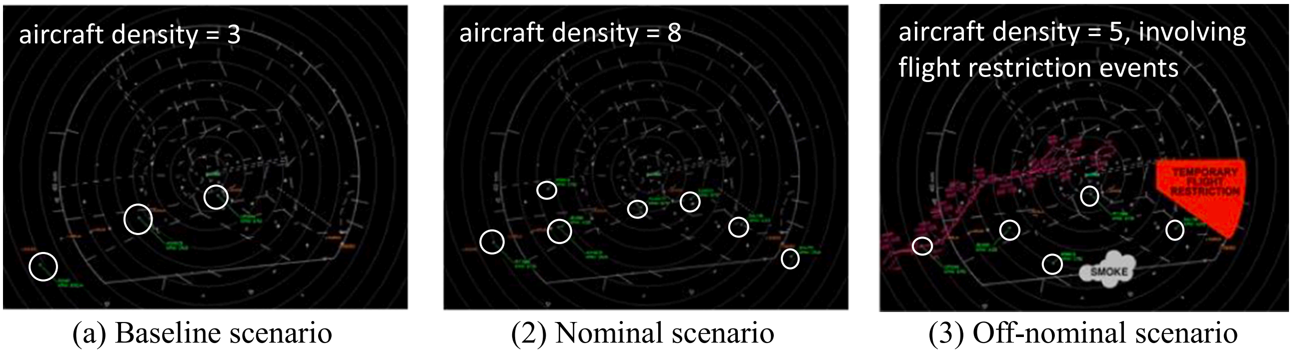


**Fig. 3.** Phoenix terminal airspace on three simulated air traffic management scenarios.

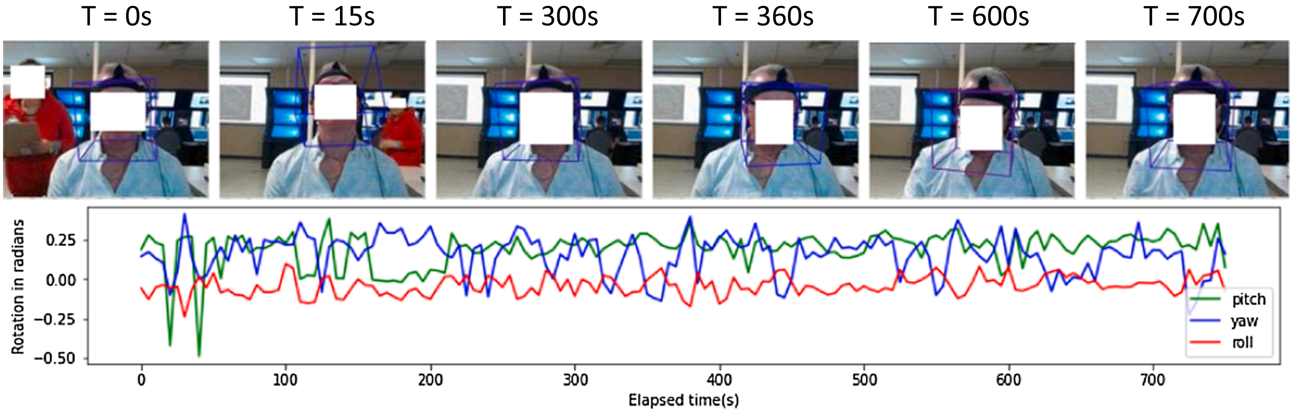


**Fig. 4.** Head pose estimation results based on the OpenFace algorithm.

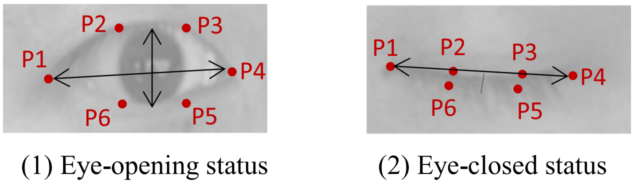


**Fig. 5.** Eye landmarks in eye-opening and closed states.

average pooling and a soft-max activation function to produce a facial expression prediction.

### 3.2.3. *Physiological behavior measures*

Physiological signals can help recognize spontaneous physiological activities and relate them to human factors (e.g., cognitive load and mental states) in air traffic control. This study collected controllers' heart rate activities using ECG equipment. By attaching two leads to the participant's left and right collar bones, the ECG equipment can measure the echoes of the heart's electrical activity in real-time during the air traffic control simulations. Specifically, the authors used the inter-beat interval (IBI), the time in milliseconds between two consecutive heartbeats, to characterize controllers' mental states. Fig. 7 shows an example

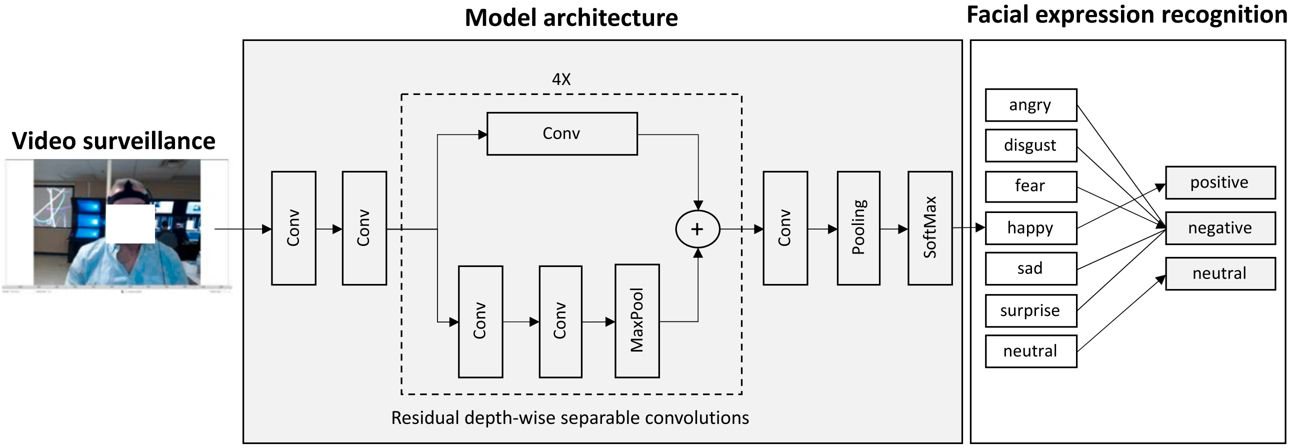


**Fig. 6.** Facial expression recognition for ATCOs.

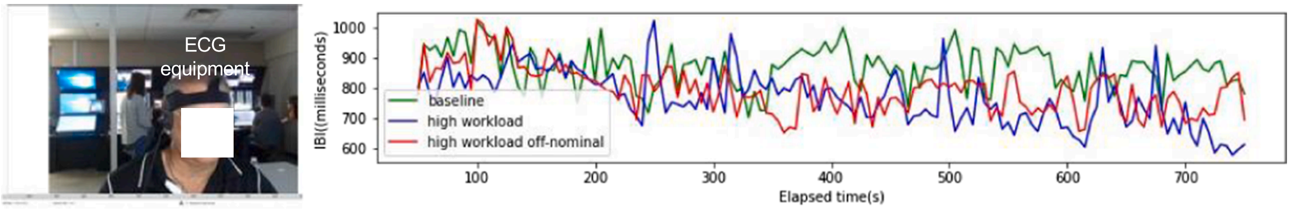


**Fig. 7.** Inter-beat interval (IBI) measures of one participant over three task scenarios.

of IBI measures for one ATCO participant across three different task scenarios. The examples show that the IBI varies across different task scenarios.

#### 3.2.4. *Mental workload measures*

Mental workload refers to the mental effort required by controllers to perform traffic control tasks. High cognitive load can negatively impact performance and increase the probability of separation errors. This study collected subjective workload measures during the simulation using a 7-point scale (from very low to very high) developed by [9]. Participants were asked to respond to probe questions as quickly as possible but were required to prioritize the air traffic management tasks per instructions. To reduce potential reporting bias, the authors gathered workload ratings three times (at 180 s, 720 s, and 1260 s) during each task scenario.

#### 3.2.5. *Performance measures for air traffic controllers*

Errors made by controllers during air traffic control operations include separation errors, controller-pilot communication errors, handover errors, and procedural errors [71]. Of these errors, separation errors are of particular concern in the aviation industry due to the expected growth in air traffic [20,64], as they can lead to LoS events between aircraft in the airspace. For this reason, this research focuses on separation errors as a measure of controllers' operational performance.

According to the mandated separation standard set by International Civil Aviation Organization (ICAO) [72], a separation error occurs when the distance between two or more aircraft is less than 5 nautical miles horizontally or 1000 feet vertically. Aircraft are considered to have sufficient separation when the horizontal and vertical separation minima with other aircraft are satisfied. If these standards are not met, the situation is classified as an LoS event. In this study, the authors used simulation tracklogs to identify separation errors made by each participant during air traffic operations. This information was then used to evaluate the performance of the proposed multimodal data analytics framework for predicting separation errors in air traffic control.

### 3.3. *Multi-step separation error prediction model*

This study proposed a multi-step separation error prediction model that combines task and behavioral factors, including air traffic complexity, visual behaviors, and physiological signals. The joint representation of multimodal features allows for capturing performance-related information within and across modalities, thereby improving the prediction model performance.

The proposed encoder-decoder LSTM network can map multimodal input sequences into multi-step output sequences, as shown in Fig. 8. LSTM networks [73] can learn long-term dependencies, which makes them well-suited for classifying and processing time series behavioral data. Within LSTM models, three gates (input gate, forget gate, and output gate) collectively control and update the information flow into and out of the cell. The input gate determines with which values to update the memory state. The forget gate determines the extent to forget the previous outputs and selects the optimal time lag for the input sequences. The output gate regulates the output values to the next hidden state based on inputs and the block's memory.

Eq. (2) – (5) represent the description of the input $i_t$, forget $f_t$ and cell activation $\widetilde{c}_t$, and output $o_t$, respectively, which enable the LSTM cell to predict the output vector:

$$f_t = \sigma\left(w_{f,x}x_t + w_{f,h}h_{t-1} + w_{c,i}c_{t-1} + b_f\right) \tag{2}$$

$$i_t = \sigma\left(w_{i,x}x_t + w_{i,h}h_t + w_{c,f}c_{t-1} + b_i\right) \tag{3}$$

$$\widetilde{c}_t = tanh\left(w_{\widetilde{c},x}x_t + w_{\widetilde{c},h}h_{t-1} + b_{\widetilde{c}}\right) \tag{4}$$

$$o_t = \sigma\left(w_{o,x}x_t + w_{o,h}h_{t-1} + b_o\right) \tag{5}$$

where $w_{f,x}$, $w_{i,x}$, $w_{\widetilde{c},x}$, and $w_{o,x}$ represent the weight coefficients, which can map the input of the hidden layer to the gates and input cell; $w_{f,h}$, $w_{i,h}$, $w_{\widetilde{c},h}$, and $w_{o,h}$ are the weight coefficients, which can connect the previous cell output state to the gates; $b_f$, $b_i$, $b_{\widetilde{c}}$, and $b_o$ are four bias vectors. $\sigma(\bullet)$ represents the standard logistics sigmoid function.

The developed model consists of two LSTMs, one serving as an encoder and the other as a decoder (see Fig. 8). The encoder maps an input sequence to a vector representation of fixed dimensionality, while the decoder is another LSTM network that uses this vector representation to produce the target sequence. The LSTM modules stack multiple layers on the encoder and the decoder to improve their ability to understand the complex representation of the time-series features at different levels. Specifically, the encoder converts the given input sequence $\mathbf{X}_i$ to a fixed-length vector, known as the context vector, which characterizes the internal representation of the input sequence. A repeat vector layer then repeats the context vector for *n*-steps (*n* is the number

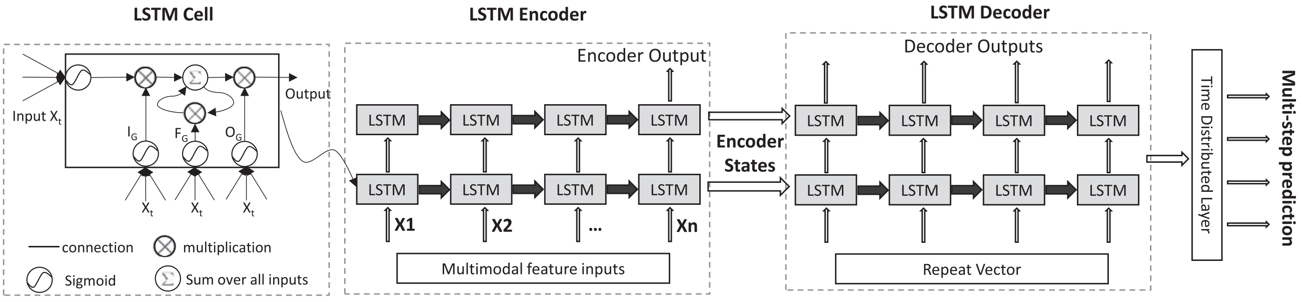


**Fig. 8.** The encoder-decoder LSTM network for predicting multi-step separation errors.

of future steps to be forecasted). The output of the repeat vector serves as the initial decoder state to predict the output sequence. The time-distributed layer is a wrapper that allows a fully connected dense layer to each time step and separates the output for each time step. In this study, the proposed model will predict future $n$-step separation errors ($n = 2$, 5, 8, and 10, corresponding to 10 s, 25 s, 40 s, and 50 s) for forecasting a sequence indicating times of controllers' separation errors.

## 4. Experiment designs

This section provides details on the experimental designs and data collection procedures employed in controller-in-the-loop simulations.

### 4.1. Experimental scenarios and procedures

The high-fidelity simulated environment used in this study was based on the Phoenix TRACON airspace, which is one of the busiest airspaces in the United States. However, unlike the Phoenix terminal in the real world, where each controller typically handles an arrival flow, this study included a second aircraft flow for aircraft and small general aircraft in the airspace to increase the overall task difficulty.

To create different levels of task complexity, the researchers adjusted the flight schedules and developed three scenarios: baseline, nominal, and off-nominal. Fig. 9 shows air traffic volume variations across three task scenarios. The baseline scenario had a relatively steady traffic density throughout the operation, with participants handling a total of 16–20 aircraft. The nominal scenario had a moderate to high traffic density, with participants controlling around 30–35 total aircraft. The off-nominal scenario had a similar traffic density to the nominal scenario but also included off-nominal events during the operation. Table 5 lists the four off-nominal events and their scripts used in the experiment, including turbulence reports, NORDO (no radio), active runway changes, and aircraft minimum fuel reported. All participants completed all three task scenarios in a randomized order.

The study protocol consisted of three within-subjects and counter-balanced ATM scenarios with different task complexities: baseline, nominal, and off-nominal, as shown in Fig. 10. The counterbalanced design, which randomizes the scenario orders of each participant, can help reduce the sequence effects of procedure variables. Each within-subject scenario took 25 mins to complete. Before the experiments, each participant received a conceptual introduction and hands-on training with simulation workstations. All participants were previously informed about the purpose of the test, its procedure, the equipment to be used, and the expected duration. During the experiment, participants were responsible for performing standard air traffic control tasks, such as resolving conflicts, issuing clearances, and providing traffic information. All pilot-controller communication was conducted via voice. Participants were asked to respond to workload probe questions by touching a button on the screen only when they had sufficient capacity during the task simulations. After the experiments, participants were encouraged to provide suggestions or comments for improving the simulations.

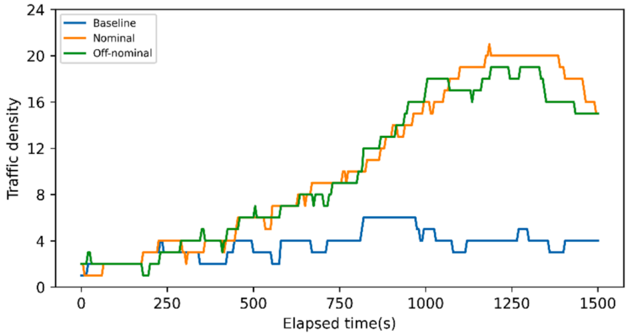


**Fig. 9.** Air traffic density variations across different task scenarios.

**Table 5**
Off-nominal events in the high workload off-nominal scenarios.

| Events | Time of events | Scripts |
|---|---|---|
| Turbulence reported | At 5mins | Pseudo-pilot script: "Approach, speed bird two eighty-one, we're experiencing moderate turbulence at one three thousand." |
| Loss of Radio Communication (NORDO) | At 10 min | Pseudo-pilot script: one aircraft does not respond to controller communication |
| Active runway switches | At 14 min | Ghost controller script: "Quartz, this is local south. We're switching to runway seven left and seven right effective immediately." |
| Minimum fuel reported | At 18 min | Pseudo pilot script: "Approach, twenty-four, minimum fuel advisory, cannot accept delayed arrival." |

### 4.2. Participants and apparatus

This study was conducted in two phases. The first phase of data collection (Phase I) took place from March 2019 to October 2019 and involved six professional participants ($N = 6$) who were retired certified air traffic controllers. The pseudo-pilots were senior students in the aviation program at Arizona State University with extensive experience in controlling aircraft using the flight simulator. All participants in this phase were male and had a professional background in air traffic control. Data collection was paused during the COVID-19 pandemic. The second data collection phase took place from July 2021 to February 2022 and involved four non-professional participants ($N = 4$) who were graduate students in ATM programs. However, visual behaviors were not available in this phase due to the use of masks, which caused facial occlusions. The institutional review board (IRB) of the institutes involved approved all of the experiments. Table 6 provides a summary of the details for both phases of data collection.

Fig. 11 shows the air traffic control simulator used for data collection, which consists of eight operational workstations for pseudo pilots and controllers. Each workstation has a radar screen, flight strips, automatic terminal information service, an external camera, and radio equipment. During the experiments, the following data were collected while controllers were dealing with different scenarios:

- **Simulator tracklogs**: The authors used the Metacraft SimSuite platform [74], which imitates FAA's standard terminal equipment, to design and simulate air traffic operations based on real traffic samples. The simulator tracklogs, updated every 5 s, recorded airspace data such as flight ID, aircraft type, and aircraft states (e.g., altitude, latitude, longitude, and speed).
- **Video data**: A high-definition webcam (Logitech C920 HD Pro Webcam) is mounted on each workstation for collecting videos of the controllers' head regions, which can be used to extract essential visual behaviors, including head poses, eyelid movements, and facial expressions. The video data collection is collected at a frequency of 30 frames per second (FPS).
- **Physiological data**: The ABM B-Alert equipment [75] continuously monitors participants' heart rates during air traffic control simulations by attaching 2 ECG leads to their collarbone areas. The heart rate data is collected at a rate of 256 samples per second.
- **Workload probes**: A touchscreen Surface tablet was used to gather participants' subjective workload ratings for traffic control tasks on a 7-point scale [9] at three different times during each scenario: 180 s, 720 s, and 1260 s.

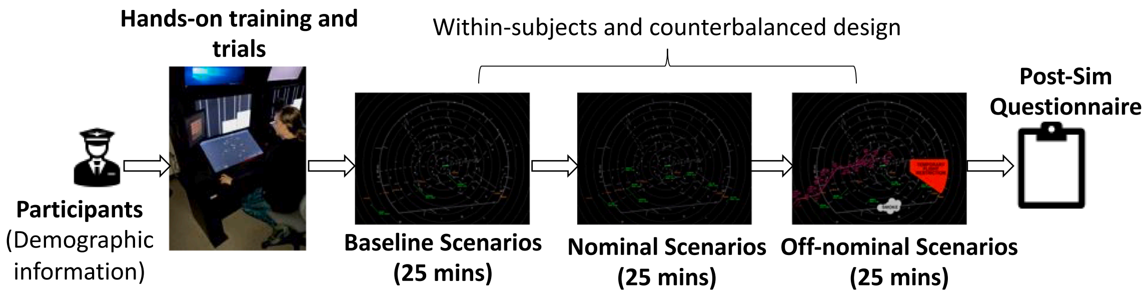

Hands-on training and trials
Within-subjects and counterbalanced design
Post-Sim Questionnaire
Participants (Demographic information)
Baseline Scenarios (25 mins)
Nominal Scenarios (25 mins)
Off-nominal Scenarios (25 mins)

**Fig. 10.** Experimental procedures of the controller-in-the-loop simulations.

**Table 6**
Summary of two data collection phases.

| Phases | Subjects | Background | Total task sessions | Total task durations | Total samples | Data types | Frequencies of data collection and designed self-reporting |
|---|---|---|---|---|---|---|---|
| Phase I | 6 | Retired ATCOs | 18 (6*3) | 450 mins | 5,418 (301*18) | Simulator tracklogs | Every 5 s in each task scenario |
| | | | | | | Workload probes | At 180 s, 720 s, and 1260 s |
| | | | | | | Video data | 30 FPS |
| | | | | | | ECG data | 256 samples per second |
| Phase II | 4 | Graduate students from the Air Traffic Management (ATM) program | 12 (4*3) | 300 mins | 3,612 (301*12) | Simulator tracklogs | Every 5 s in each task scenario |
| | | | | | | Workload probes | At 180 s, 720 s, and 1260 s in each task scenario |
| | | | | | | ECG data | 256 samples per second |
| | | | | | | Note: Phase II did not collect the visual data due to the mask coverings. | |

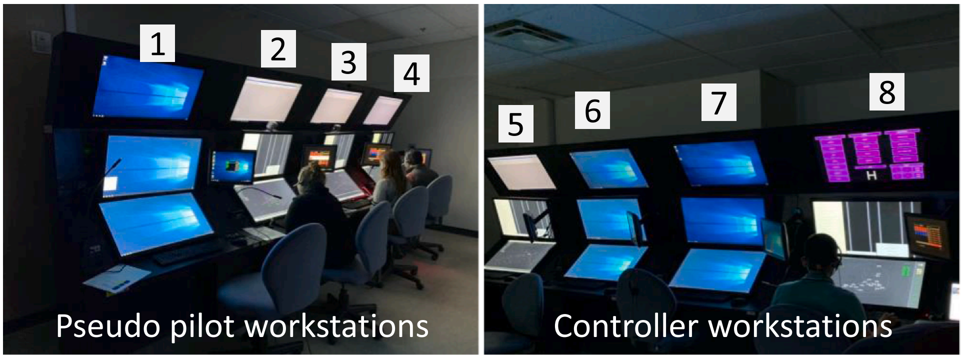


**Fig. 11.** The Phoenix TRACON simulator (eight operational workstations) for pseudo pilots and controllers.

## 5. Experimental results and analysis

This section presented the results of a factor analysis of the proposed multi-factorial model and evaluated the performance of the developed encoder-decoder LSTM prediction model using different features. The authors also compared the developed model with three baseline deep learning methods for predicting multi-step separation errors.

### *5.1. Factor analysis of the multi-factorial model*

This subsection analyzed the proposed multi-factorial model using all measures from the six professional ATCOs in Phase I who did not wear facial masks. In the discussion section, the authors will further compare the results of this analysis with those obtained using data collected in Phase II, which included four non-professional ATCOs.

#### *5.1.1. Analysis of variance for measures across air traffic management scenarios*

This study examined the descriptive statistics for all measures, including traffic density, mental workload, separation error, visual behaviors, and physiological signals, for three task scenarios. Specifically, the workload ratings were averaged across the 25-minute task scenarios (three measurements) to facilitate comparisons between the different task demands. The results (Fig. 12) suggest that the traffic density and workload ratings varied as expected across the three task scenarios. The traffic density was the lowest in the baseline scenario (scenario 1), while it was much higher in the nominal and off-nominal scenarios (scenarios 2 and 3). The baseline scenario was rated as having the lowest workload ratings, while participants reported a higher mental workload in the nominal scenario as the traffic density increased. However, the off-nominal scenario, which included off-nominal events such as turbulence reported, NORDO (no radio), active runway changes, and aircraft minimum fuel reported, had slightly higher workload demands, as reported by the participants.

When comparing the number of separation errors made in the three task scenarios, the authors found that the baseline scenario has fewer separation errors than the nominal and off-nominal scenarios. Nonetheless, the nominal and off-nominal scenarios have similar separation

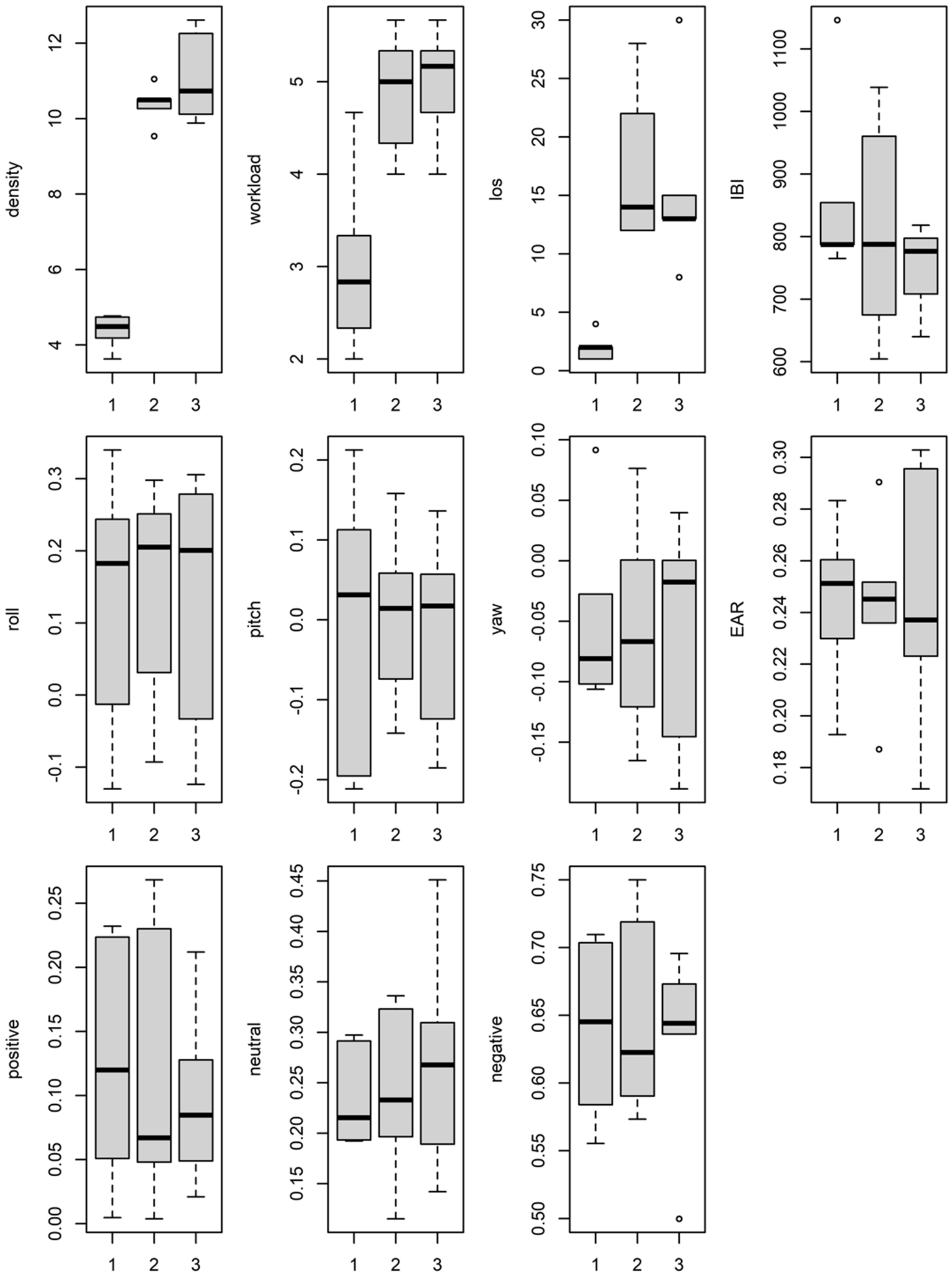


**Fig. 12.** Boxplots for all measures in three task scenarios (1 = baseline scenario, 2 = nominal scenario, 3 = off-nominal-scenario).

errors, although the off-nominal scenario included off-nominal events. This finding suggests that the introduction of off-nominal events did not significantly impact the number of separation errors made in the scenario.

To better understand the statistical effects of all measures, the authors conducted a repeated measures analysis of variance (ANOVA) [76] for each task scenario. For each analysis, Mauchly's sphericity was tested to verify whether the variances of the differences between all possible pairs of within-subject conditions are equal. If sphericity was violated, Greenhouse–Geisser (GG) correction was applied for adjusted p-values. The generalized effect size $\eta_G^2$, which measures the degree of association between the effect and the variables, was also reported.

The results in Table 7 show that significant differences in traffic density ($p < .001$, $\eta_G^2 = 0.949$), mental workload ($p < .001$, $\eta_G^2 = 0.65$), and separation error ($p < .01$, $\eta_G^2 = 0.615$) exist across the three task scenarios. The $\eta_G^2$ values also indicate that these differences have large

**Table 7**
Statistical analysis of measures in three air traffic management scenarios.

| Measures | Task scenarios | | | Repeated measures ANOVA | | |
|---|---|---|---|---|---|---|
| | Baseline | Nominal | Off-nominal | $F$ | $p$ | $\eta_G^2$ |
| Density | 4.382(0.450) | 10.394(0.494) | 11.057(1.144) | 113.403 | 3.4e-05$^{***}$ | 0.949 |
| Workload | 3.000(0.943) | 4.889(0.621) | 5.000(0.596) | 50.328 | 0.0006$^{***}$ | 0.65 |
| LoS | 2.000(1.095) | 17.000(6.573) | 15.333(7.554) | 10.340 | 0.004$^{**}$ | 0.615 |
| IBI | 868.012(159.025) | 813.134(184.264) | 744.904(93.266) | 3.254 | 0.145 | 0.331 |
| Yaw | −0.051(0.076) | −0.057(0.088) | −0.055(0.091) | 0.048 | 0.854 | 0.001 |
| Pitch | −0.003(0.170) | 0.005(0.107) | −0.014(0.120) | 0.228 | 0.8 | 0.004 |
| Roll | 0.134(0.174) | 0.150(0.152) | 0.138(0.175) | 0.336 | 0.722 | 0.002 |
| EAR | 0.245(0.031) | 0.243(0.033) | 0.245(0.049) | 0.042 | 0.959 | 0.0008 |
| Positive | 0.125(0.092) | 0.114(0.108) | 0.097(0.070) | 1.047 | 0.386 | 0.02 |
| Negative | 0.640(0.062) | 0.646(0.072) | 0.632(0.069) | 0.094 | 0.911 | 0.009 |
| Neutral | 0.234(0.048) | 0.240(0.083) | 0.271(0.110) | 0.606 | 0.564 | 0.044 |

Note: The standard deviations are given in parentheses. *. $p < .05$, **. $p < .01$, ***. $p < .001$. $\eta_G^2$ is the generalized effect size. Small effect size: $\eta_G^2 \leq 0.01$; Medium effect size: $0.01 < \eta_G^2 \leq 0.14$; and large effect size: $\eta_G^2 > 0.14$.

effect sizes. However, the results show that there are no significant differences in visual behaviors (i.e., head pose, eyelid movement, and facial expressions) and physiological behaviors (i.e., IBI) between the three scenarios. This finding is consistent with previous research [77] and suggests that these behavioral measures may not be directly related to task complexity but instead reflect participants' mental states, such as their levels of attention and vigilance. While abnormal mental conditions can greatly impact the operational performance of ATCOs, the lack of a direct relationship between behavioral indicators and separation errors highlights the need to further explore the complex influence of multiple human and task factors on operational processes. This may involve in-depth modeling of the relationships between these factors and separation errors.

The authors also conducted *post-hoc* tests following the Bonferroni procedure [78], which adjusts p-values and eliminates multiple spurious positives, to analyze the measure differences among three task scenarios. The results of pairwise comparisons (shown in Fig. 13) revealed that traffic density and mental workload were significantly lower in the baseline scenario than in the nominal and off-nominal scenarios. However, the traffic density ($p = 1.0$) and workload ratings ($p = 0.525$) were not significantly different between the nominal and off-nominal scenarios. Similarly, the number of separation errors made in the baseline scenario was significantly lower than in the nominal and off-nominal scenarios, but there were no significant differences between the nominal and off-nominal scenarios ($p = 1.0$). These results suggest that traffic density is one of the main influences on workload and separation errors in task scenarios.

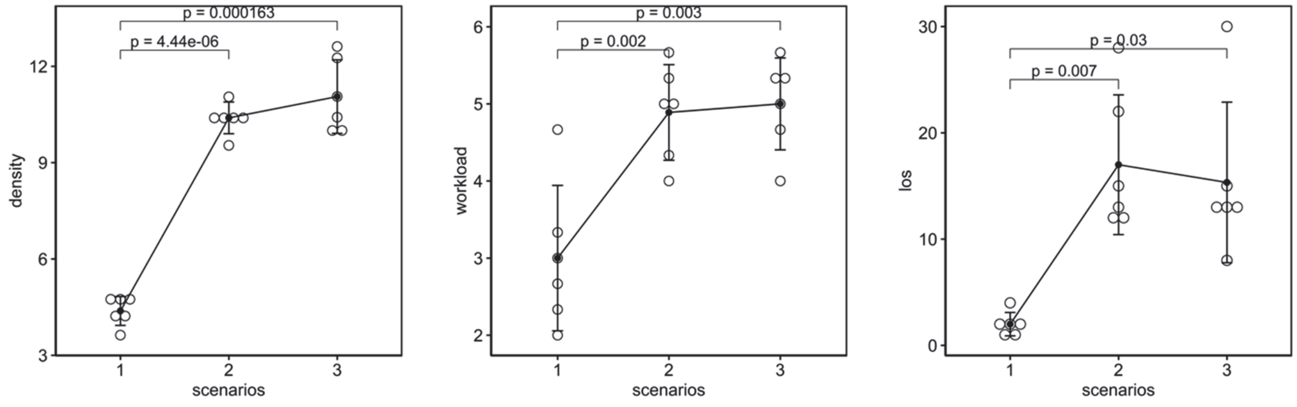
### 5.1.2. Correlation between task complexity, workload, and separation errors

To investigate the inter-relationships between task complexity, workload, and separation errors, the authors conducted a Spearman correlation analysis [79] with a 95% confidence interval. The correlation coefficient $r$, which ranges from −1 to 1, indicates the relationship between two variables. A high positive $r$ represents a proportional relationship, while a high negative $r$ indicates an inverse relationship.

As shown in Fig. 14, the traffic density was significantly correlated with the workload ($r = 0.56, p < .05$). Similarly, the traffic density ($r = 0.64$, $p < .01$) and mental workload ratings ($r = 0.67$, $p < .01$) were significantly associated with separation errors. These results suggest that the changes in the task complexity (traffic density in the designed scenarios) could lead to variations in mental workload and subsequently impact ATCOs' operational performance. Overall, the factor analysis and correlation analysis results support the importance of considering both task complexity and workload in predicting separation errors made by air traffic controllers. The developed encoder-decoder LSTM model, which considers multiple factors, including task complexity, workload, and behavioral indicators, has the potential to improve the accuracy of separation error prediction in air traffic control operations.

### 5.1.3. Correlation between facial and physiological features

To determine the correlation between facial and physiological features, the authors analyzed the means and standard deviations of measures, including inter-beat intervals (IBI), head pose (roll, pitch, and yaw), eye aspect ratio (EAR), and facial expressions (positive, neutral, and negative), as shown in Fig. 15. The results of the Spearman

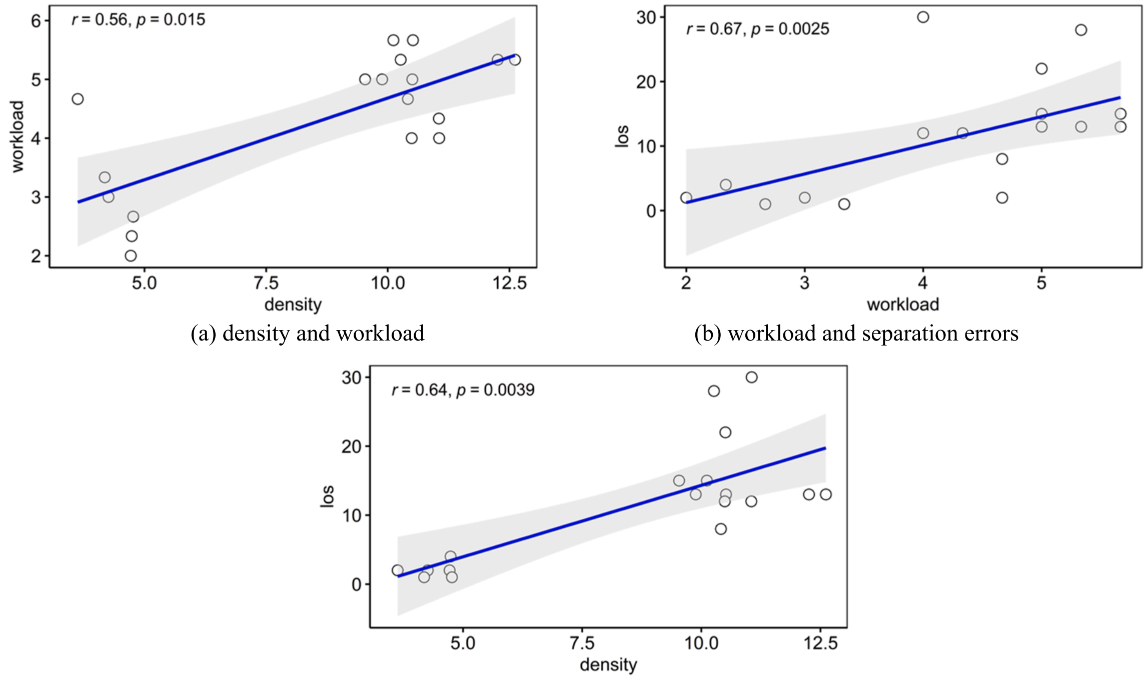


(a) density

(b) workload

(c) separation errors

**Fig. 13.** *Post hoc* tests for measures in three task scenarios (1 = baseline scenario, 2 = nominal scenario, 3 = off-nominal-scenario).

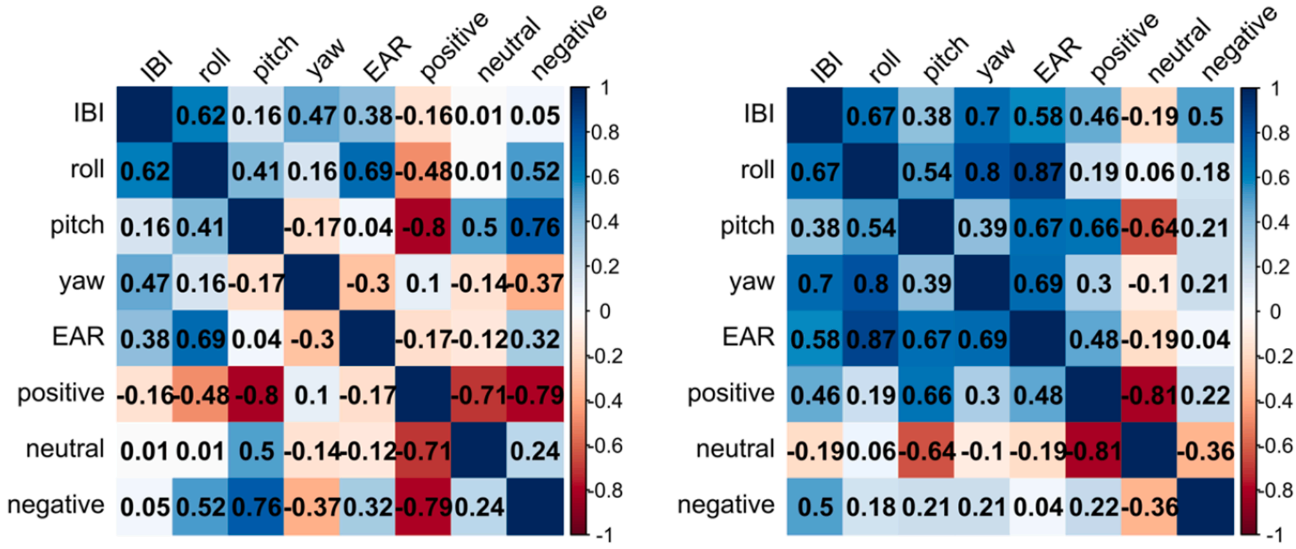


(a) density and workload

(b) workload and separation errors

(c) density and separation errors

**Fig. 14.** Correlation analysis of the traffic density, workload, and separation errors.



(a) Means of measures

(b) Standard deviations of measures

**Fig. 15.** Correlation analysis of facial and physiological features. Note: The color bar shows the correlations range in magnitude from −1.00 to 1.00.

correlation revealed a moderate relationship between facial and physiological features. A high positive relationship ($r > 0.7$) was found between mean pitch angles and negative facial expressions and between standard deviations of EAR and roll angles. In contrast, a moderate negative relationship ($r <$ -0.5) was found between standard deviations of pitch angles and neutral facial expressions, and a high negative relationship ($r <$ -0.7) was identified between mean pitch and positive facial expressions. Furthermore, a moderate positive relationship ($r > 0.5$) was found between mean IBI and roll, between standard deviations of IBI and roll angles, between standard deviations of IBI and yaw angles, between standard deviations of IBI and EAR, between standard deviations of IBI and negative facial expressions, between standard deviations of the EAR and head pose (i.e., pitch and yaw), between mean head pose (i.e., roll and pitch) and negative facial expressions. These results suggest that facial and physiological features are related and can be used together to predict separation errors. The combination of multiple features could provide a more robust prediction model compared to using a single feature.

### 5.2. *Performance analysis of the multi-step separation error prediction model*

As the visual data in Phase II is unavailable, this study used all 18 task session data (three scenarios for each participant) collected from six professional ATCOs to evaluate the developed multi-step separation error prediction model.

### 5.2.1. Implementation details

This study collected raw data from multiple sources at different sampling rates. To synchronize the data, the authors merged raw data with a 5 s synchronized timestep (informally, “step”) based on the update period of the Metacraft simulation system, which records air traffic information every 5 s. Using the simulation tracklogs, a separation status (LoS or non-LoS) was assigned to each timestep of each participant. Each participant had approximately 301 steps (including the step at the beginning of each experiment) across one ATM scenario ((25 min × 60 s) / 5 s + 1). The resulting dataset consists of 5,418 steps labeled with separation status and contains nine features, including task density, head pose (pitch, roll, and yaw), eyelid movement, facial expressions (happy, neutral, and negative), as well as IBI data. In particular, the dataset includes 33.52% positive samples or timesteps with separation errors.

This study used *k*-fold cross-validation [80], a common method for evaluating machine learning models on a limited amount of data. In this method, the data is divided into *k* subsets, and the proposed model is trained and evaluated *k* times, with a different subset used as the test set each time. The authors set *k* as 10, meaning that 90% of the data was used for training and 10% was used for testing. In each fold, a 15% subset of the training data was used for validation. The performance of the model was then averaged over the 10 folds.

To preprocess the data, the authors used the min–max scaling to rescale the measures in the dataset. The LSTM layer of the encoder-decoder model contains 64 hidden units. The loss function used is the mean squared error (MSE), and the activation function is the ReLU. During training, the model was optimized using the Adam optimizer with a learning rate of 0.001 and a decay rate was 0.9. The batch size was set to 8, and the maximum number of epochs was 200 with early stopping implemented. The length of the historical input data (i.e., observation length) was a hyperparameter that was optimized by selecting the parameter that resulted in the highest accuracy on the validation dataset. In this study, the observation length was set to 15 steps, as this setting resulted in the highest prediction accuracy. The model did not show any significant improvement in performance with longer observation lengths.

Fig. 16 shows the model performance of the loss function using (a) a single feature (i.e., traffic density) and (b) all features. The loss curves decrease as the number of training epochs increases. The loss functions are similar in both the training and validation phases, indicating that the network has generalized well. Additionally, the final losses of the model trained with all features are lower than that of a single feature, indicating that using all features results in better prediction of separation errors.

### 5.2.2. Model performance evaluation metrics

This study evaluated the prediction performance of the proposed encoder-decoder LSTM model through two types of metrics: (1) point-level metrics, including *precision*, *recall*, and *F*1-score, and (2) sequence-level metrics, including alignment accuracy and sequence similarity.

The authors define that true positive (*TP*) is the number of timesteps correctly predicted as separation errors, false positive (*FP*) is the number of timesteps incorrectly identified as separation errors, and false-negative (*FN*) is the number of timesteps incorrectly identified as non-separation errors. The *precision*, *recall*, and *F1*-score are defined as follows:

$$Precision = TP \,/\, (TP + FP) \tag{6}$$

$$Recall = TP \,/\, (TP + FN) \tag{7}$$

$$F1 = 2 \times Precision \times Recall \,/\, (Precision + Recall) \tag{8}$$

Alignment accuracy and sequence similarity are metrics commonly used to evaluate the model performance that aligns two sequences, particularly in protein and DNA sequence analysis [81]. This study used these two metrics to assess the model performance of the proposed method for predicting separation errors at various steps following a given time. The definitions of alignment accuracy and sequence similarity are as follows:

$$Alignmentaccuracy = \frac{M_c}{M} \tag{9}$$

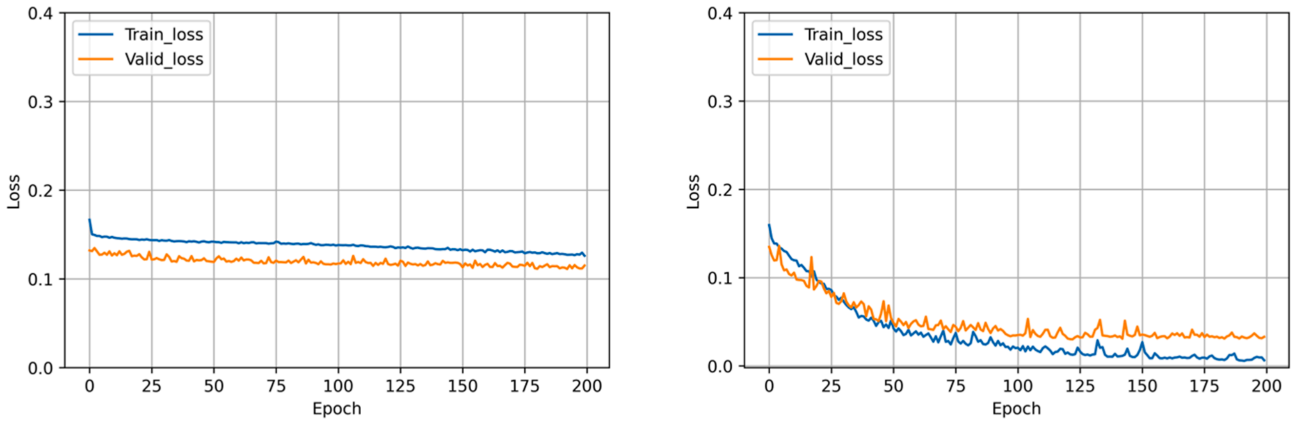
$$Sequencesimilarity = \frac{1}{M}\sum_{i=1}^{M}\frac{L_c}{L} \tag{10}$$

where $M_c$ is the number of correct sequence predictions, $M$ is the total number of sequence predictions, $L_c$ is the length of the longest common subsequence for ground-truth and predicted sequences, and $L$ is the length of the ground-truth sequence.

### 5.2.3. Prediction performance with different features

This study compared the prediction performance of multiple input features, including task features (e.g., traffic density), behavioral features (e.g., head pose, eyelid movement, facial expression, and IBI), and hybrid features of both task and behavioral data. Each type of input feature was used to create a separate prediction model.

Fig. 17 presents the *precision*, *recall*, and *F*1-score performance of the proposed method when using single input features. The results indicate that the head-pose feature performs better than any other single feature, with 89.38% *precision,* 88.12% *recall*, and 88.75% *F*1 score in predicting



(a) Single feature (i.e., traffic density)

(b) All features

**Fig. 16.** The loss curves of model performance from a randomly selected fold.

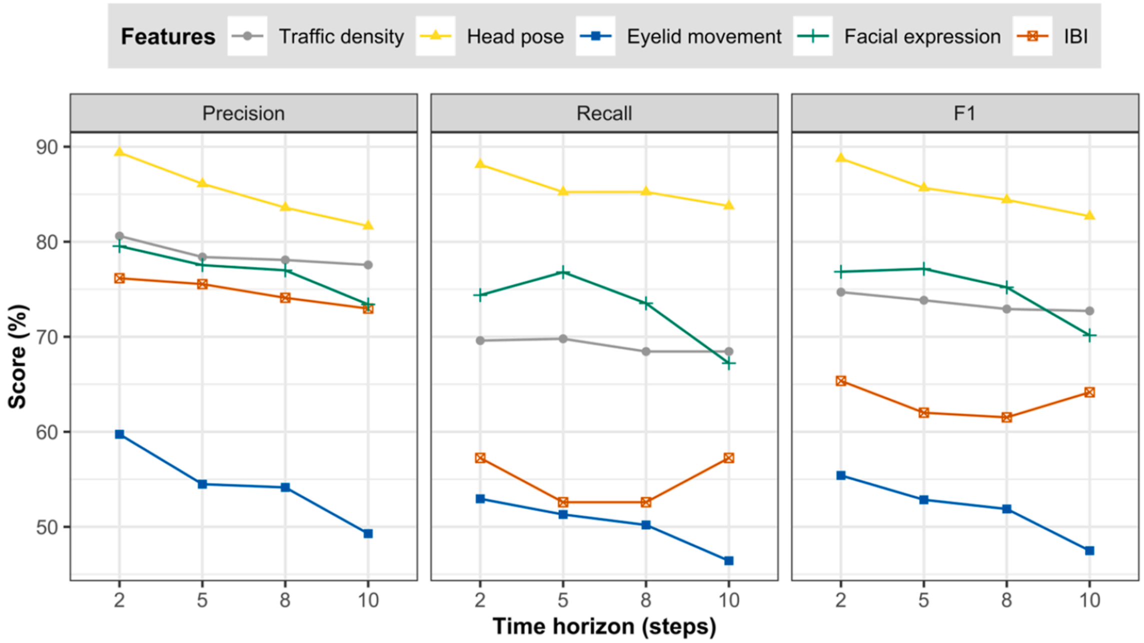


**Fig. 17.** Precision, recall, and F1-score performance of the proposed method based on single features.

2-step separation errors. The head poses of controllers seemed to be more relevant in indicating their operational performance, as they showed their visual attention between multiple. This highlights the importance of the head-pose feature and suggests that visual behaviors of ATCOs can serve as indicators of errors in addition to task features and physiological signals.

Meanwhile, the results of this study revealed that the eyelid movement feature had relatively lower performance in predicting separation errors, with a *precision* of 59.74%, a *recall* of 52.95%, and an *F*1 score of 56.14% in 2-step separation error prediction. One possible reason for this lower performance is that some ATCO participants wore glasses, which can affect the accuracy and clarity of the extracted eyelid features due to reflections. Despite this, the proposed encoder-decoder LSTM method showed relatively stable performance across different time horizons, indicating that this method can capture complex patterns and reduce accumulation errors in long-term forecasts. This suggests that the method effectively addresses the challenges of predicting separation errors over longer periods.

Table 8 presents the alignment accuracy and sequence similarity performance of the proposed method using single features. The results revealed that the head-pose feature is the most informative (alignment accuracy = 82.12%, sequence similarity = 87.61%), while the eyelid movement feature is the least informative (alignment accuracy = 65.03%, sequence similarity = 70.37%) when only integrating a single feature to predict 2-step separation errors. This conclusion confirmed the results of the point-level performance analysis. The results in Table 8 also indicated that the facial expression and IBI could achieve comparable performance and provide relatively high alignment accuracy (over 70% alignment accuracy in 2-step prediction) in the short-term prediction of controllers' separation errors. The results of this study showed that predicting separation errors becomes more complex as the number of steps being forecasted increases. For example, the head-pose feature had nearly a 10% higher alignment accuracy when predicting 2-step separation errors compared to 10-step forecasting. This illustrates that many-step forecasting is more challenging than 2-step forecasting. However, the sequence similarity score did not decrease significantly, suggesting that the proposed encoder-decoder LSTM network could accurately capture the trends in the data. This observation suggests that while the task becomes more difficult as the number of steps being forecasted increases, the proposed model can handle this increased complexity effectively.

This study also investigated the performance of combining task and behavioral features as inputs. Fig. 18 shows the *precision*, *recall*, and *F*1-score performance of the proposed method using hybrid features. Table 9 presents the alignment accuracy and sequence similarity performance of the proposed method using hybrid features. These results suggest that multimodal features could improve prediction performance compared to single features. This may be because integrating different features can provide complementary and cross-modality information between modalities. In particular, the model that combines all visual features (i.e., head pose, eyelid movement, and facial expression) and task contexts performed better or comparably to the model integrating

**Table 8**
Alignment accuracy and sequence similarity of the proposed method based on single features.

| Features | 2 steps | | 5 steps | | 8 steps | | 10 steps | |
|---|---|---|---|---|---|---|---|---|
| | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity |
| Traffic density | 81.47 | 83.33 | 77.61 | 82.62 | 73.80 | 81.26 | 70.51 | 81.32 |
| Head pose | **82.12** | **87.61** | **78.79** | **87.55** | **74.75** | **86.73** | **71.70** | **87.09** |
| Eyelid movement | 65.03 | 70.37 | 57.77 | 67.33 | 53.83 | 63.76 | 50.21 | 61.80 |
| Facial expression | 75.62 | 79.45 | 70.62 | 78.72 | 66.89 | 77.93 | 63.82 | 78.10 |
| IBI | 72.75 | 74.58 | 68.07 | 74.81 | 64.41 | 75.68 | 62.87 | 77.33 |

Note: The highest metric values are in bold. All metric values are in percentage.

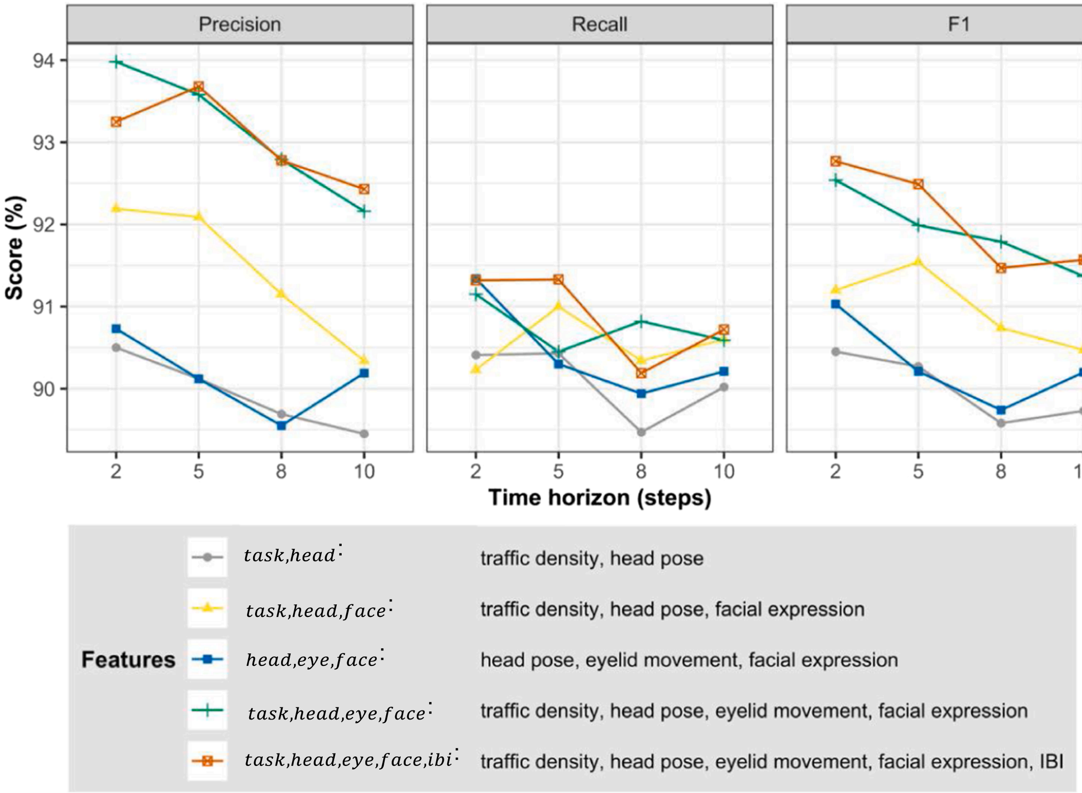


**Fig. 18.** Precision, recall, and F1-score performance of the proposed method based on hybrid features.

**Table 9**
Alignment accuracy and sequence similarity of the proposed method based on hybrid features.

| Features | 2 steps | | 5 steps | | 8 steps | | 10 steps | |
|---|---|---|---|---|---|---|---|---|
| | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity |
| $F_{task,\ head}$ | 88.64 | 90.13 | 84.90 | 90.92 | 82.19 | 90.80 | 79.58 | 90.72 |
| $F_{task,\ head,\ face}$ | 90.60 | 91.81 | 86.96 | 91.82 | 84.64 | 91.41 | 81.34 | 91.37 |
| $F_{head,\ eye,\ face}$ | 89.28 | 90.34 | 84.44 | 90.03 | 81.37 | 89.78 | 78.80 | 89.40 |
| $F_{task,\ head,\ eye,\ face}$ | **91.98** | **93.37** | **88.15** | 93.30 | 84.10 | 93.38 | 82.38 | 92.98 |
| $F_{task,\ head,\ eye,\ face,\ ibi}$ | 91.90 | 93.21 | 88.01 | **93.71** | **86.58** | **93.76** | **84.04** | **93.60** |

Note: The highest metric values are in bold. All metric values are in percentage.

all features. For example, $F_{task,\ head,\ eye,\ face}$ model achieved 93.98% *precision*, 91.15% *recall*, 92.54% $F1$ score, 91.98% alignment accuracy, and 93.37% sequence similarity score in predicting 2-step separation errors. This suggests that these three visual behavioral indicators are sufficient and necessary for predicting controllers' separation errors. This study also found that as the prediction time step increases, the prediction ability of all models decreases gradually. This reflects the difficulties of cumulative errors faced by multi-step forecasting models.

### 5.2.4. Comparisons with baseline models

The authors compared the proposed encoder-decoder LSTM network with three baseline deep learning models (CNN, classic LSTM, and GRU) for predicting multi-step separation errors using the same dataset. CNN is a deep neural network commonly used for image and video processing, which can also be applied for time series classification and forecasting. LSTM and GRU are variations of the classical RNN model and are often used to handle sequential learning tasks. Table 10 lists the configurations of these models that were compared. In particular, the authors set the same training parameters (such as learning rate, decay rate, and batch size) for each model and ran each model ten times on the collected dataset.

**Table 10**
Hyperparameters of three baseline models.

| Models | Structures (#layer) | Parameters |
|---|---|---|
| CNN | Conv1D – Max pooling – Flatten – Dense – Output | Conv1D: Filters = 64, kernel size = 3<br>Max pooling: pool size = 2<br>Dense: 64 |
| LSTM | LSTM – Dropout – Dense – Output | LSTM: units = 64<br>Dropout: rate = 0.5<br>Dense: units = 64 |
| GRU | GRU – Dropout – Dense – Output | LSTM: units = 64<br>Dropout: rate = 0.5<br>Dense: unites = 64 |

Fig. 19 illustrates the *precision*, *recall*, and $F1$-score performance of the different methods. The results indicate that the proposed method has the best performance, with 93.98% *precision*, 91.15% *recall*, and 92.54% $F1$ score in 2-step separation error prediction. Table 11 shows the alignment accuracy and sequence similarity performance of the different methods. The performance rank of each method is as follows: Proposed > LSTM > GRU > CNN by comparing their performance metrics on

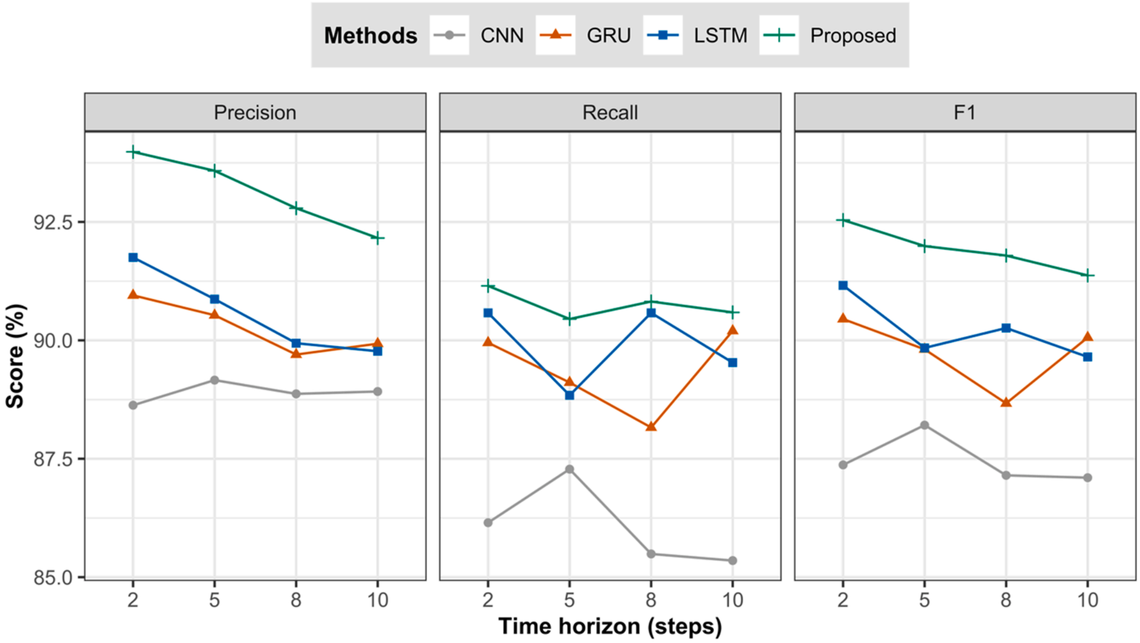


**Fig. 19.** Precision, recall, and F1-score performance of the different methods.

**Table 11**
Alignment accuracy and sequence similarity performance of the different methods.

| Methods | 2 steps | | 5 steps | | 8 steps | | 10 steps | |
|---|---|---|---|---|---|---|---|---|
| | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity | alignment accuracy | sequence similarity |
| CNN | 87.88 | 90.57 | 84.06 | 89.17 | 77.74 | 86.88 | 74.17 | 88.62 |
| LSTM | 89.60 | 91.81 | 85.14 | 91.87 | 82.92 | 91.06 | 79.24 | 90.81 |
| GRU | 88.98 | 90.64 | 84.38 | 90.55 | 81.78 | 90.58 | 78.91 | 90.32 |
| Proposed | **91.98** | **93.37** | **88.15** | **93.30** | **84.10** | **93.38** | **82.38** | **92.98** |

Note: The highest metric values are in bold. All metric values are in percentage.

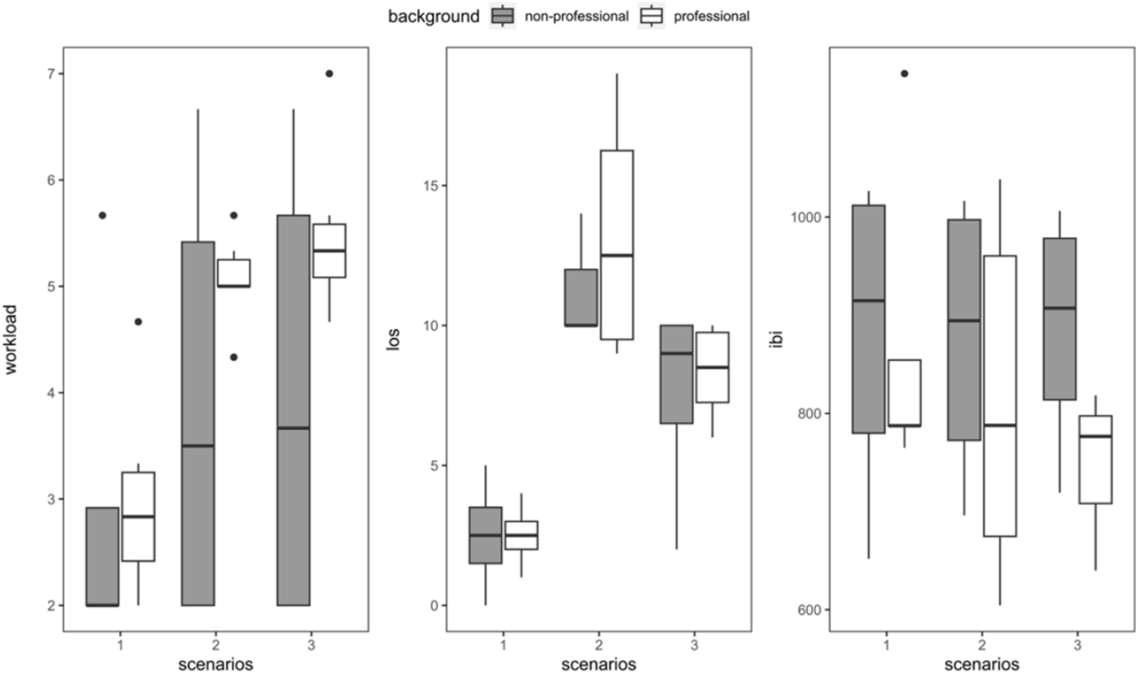


**Fig. 20.** Comparisons between the professionals (retired ATCOs) and non-professionals (graduate students) in three task scenarios.

different prediction steps. Moreover, as the prediction time step increases, our model has a greater performance advantage over the baseline models in the experiments. The results of the experiments show that the proposed model has the lowest prediction error compared to the baseline methods in both short-term and long-term time-step prediction, indicating that the proposed method can effectively learn deep representations and long-term temporal dependency features of ATCOs' separation errors. The proposed model provides a valuable reference for developing long-term prediction systems based on multivariate time series data.

## 6. Discussion

This section compared the Phase I and Phase II data collection. The authors also demonstrated the practical implications of the proposed method and explained the study's limitations for future improvements.

### *6.1. Comparisons of Phase I and Phase II data collection*

In the study, the authors compared the experimental measures obtained during Phase I and II data collection. As the visual behaviors were not available in Phase II due to facial occlusions, the measures used for comparison included mental workload, separation errors, and inter-beat intervals (IBI). The results, shown in Fig. 20, reveal that both professional (retired air traffic controllers) and non-professional (graduate students) participants perceived an increase in mental workload as the task complexity increased. However, the non-professional participants had lower reported workload ratings with a larger variance than the professional group. This may be because non-professional participants tend to optimistically assess the situational complexity, task demands, and their own cognitive capacities [82].

Additionally, the results showed that non-professional participants had comparative or relatively higher separation errors. Analysis of the aircraft trajectories revealed that non-professional participants tended to violate procedural requirements, such as height level and maximum speed limits, that are necessary for maintaining separation standards. In contrast, professional participants made fewer procedural errors during task scenarios. This suggests that future work should develop more comprehensive operation measures, particularly for non-professional participants. Moreover, non-professional air traffic controllers had higher IBI than professional air traffic controllers. This finding is consistent with Yao et al. [83], which observed that heart rates were lower for expert subjects than for novices in a flight simulator. These observations support monitoring the mental efforts of ATCOs using objective measurements.

### *6.2. Practical implications*

The research presented in this study has several theoretical and practical implications. One of the main benefits of the video-based monitoring method used in this study is that it is non-intrusive and relatively low-cost compared to expensive physiological sensors. In addition, this study used computer vision algorithms to extract multiple visual behaviors from real-time surveillance videos that are indicative of controllers' mental states. The study also compared the performance of different behavioral cues, including visual and physiological behaviors, in predicting controllers' separation errors, which can help researchers select the appropriate features for similar tasks. The results of this study also highlight the importance of using multiple behavioral measures for improved prediction accuracy.

Furthermore, the encoder-decoder LSTM network proposed in this work can use multiple features to predict separation errors in air traffic control tasks in advance. This prediction can provide timely alerts to controllers, their supervisors, or automated intelligence in airspace management systems to improve air traffic control performance, avoid separation errors, and optimize human-automation interactions. The results of this study can also be used to create a personalized training system and provide more feedback to support learners in achieving safe and efficient air traffic management performance.

### *6.3. Limitations and future work*

The proposed method has several limitations that can be addressed in future work. One limitation is the small sample size, which may impact the statistical significance of the experimental measures. To address this, future work should recruit more air traffic controllers for the experiments. Despite the limited sample size, this study still provides valuable insights into how multimodal behavioral cues can be used to infer separation errors in air traffic control tasks. The study also identified a comprehensive list of visual behavioral activities related to ATCOs' separation errors and characterized the contributions of different behavioral indicators in predicting their operational performance. This characterization can also help guide the use of specific behavioral indicators in other tasks involving the comprehension of separation errors. In addition, the high performance of the proposed encoder-decoder LSTM network suggests its potential for predicting separation errors over a longer time horizon than baseline deep learning models.

Another limitation is the task complexity design in this study, which relied on adjusting air traffic scenarios for traffic density and off-nominal scenarios. In future work, it would be beneficial to incorporate additional task complexity indicators, such as aircraft behaviors and airspace conditions, to reflect real-world air traffic control tasks more accurately. Additionally, the experiments in this study were conducted in simulated environments rather than in actual air traffic control centers. While simulations offer a more controlled environment, it is important to note that participants may have a sense of psychological comfort because they do not face the consequences of their separation errors. To address this, future studies should consider collecting data from actual air traffic control centers to obtain more realistic results.

## 7. Conclusions

Valid and reliable methods of assessing how human factors influence and indicate controllers' separation errors are critical in reducing the risk of incidents and accidents in air traffic control. This paper presented a controller-centered method for inferring ATCOs' separation errors during simulated air traffic control tasks. Our method involves developing a multi-factorial model to understand the relationships between task complexity, behavioral activity, separation errors, and cognitive load. Then, a multimodal data sensing framework simultaneously captured task information, visual behaviors, physiological indicator, and subjective workload from a high-fidelity air traffic control simulator. These visual behavioral features include head pose, eyelid movement, and facial expressions. Finally, the developed encoder-decoder LSTM network could predict separation errors by integrating multimodal features.

The authors verified the multi-factorial model using data collected from controller-in-the-loop experiments involving 18 sessions with different scenarios and six human subjects. The factor analysis of measures showed that the traffic density was significantly correlated with the mental workload ($r = 0.56$, $p < .05$). Additionally, both traffic density ($r = 0.64$, $p < .01$) and mental workload ratings ($r = 0.67$, $p < .01$) were significantly associated with separation errors. These results demonstrate that changes in task complexity can impact cognitive load and ultimately affect controllers' separation errors.

The proposed controller-centered method can help controllers prevent collisions between aircraft and proactively generate time-ahead alerts for potential separation errors by inferring the context of task information and behavioral activities. By combining all visual behaviors and task features, the developed method was able to achieve alignment accuracies of 91.98%, 88.15%, 84.10%, and 82.38% in predicting 2, 5, 8, and 10-step-ahead (corresponding to 10 s, 25 s, 40 s, and 50 s in this study) separation errors, respectively. These results demonstrate higher or comparable performance compared to other models with different hybrid features. The developed encoder-decoder LSTM model outperformed baseline deep learning models (including CNN, LSTM, and GRU) in predicting controllers' separation errors.

While the proposed method has some limitations, the authors have suggested ways to improve it in the future. One limitation is that the small number of subjects may not allow us to generalize the results to the entire population of ATCOs. In future work, it will be important to recruit more participants to address this issue. However, the current results are still useful as they show the potential of using visual human behavior analysis for this type of research. Additionally, future work should incorporate more detailed features, such as traffic flow complexity, to better characterize task demands. Finally, the authors plan to conduct experiments in real ATM environments and collect more realistic data to model controllers' operational performance.

## Declaration of Competing Interest

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

## Data availability

Data will be made available on request.

## Acknowledgments

The research was supported by funds from NASA University Leadership Initiative program (Contract No. NNX17AJ86A, Project Officer: Dr. Anupa Bajwa, Program coordinator: Koushik Datta, Principal Investigator: Dr. Yongming Liu, Co-PI: Dr. Pingbo Tang), and Bradford and Diane Smith Graduate Fellowship. The support is gratefully acknowledged. The authors would like to thank all the participants for their generous collaboration.

## References

[1] A. Nunes, A.F. Kramer, Experience-based mitigation of age-related performance declines: Evidence from air traffic control, J. Exp. Psychol. Appl. 15 (2009) 12–24, https://doi.org/10.1037/A0014947.

[2] German Federal Bureau of Aircraft Accidents Investigation, Investigation report, 2004. https://reports.aviation-safety.net/2002/20020701-1_B752_A9C-DHL_T154_RA-85816.pdf (accessed February 25, 2022).

[3] Aircraft and Railway Accidents Investigation Commission, Aircraft accident investigation report, 2002. https://www.mlit.go.jp/jtsb/eng-air_report/JA8904.pdf (accessed February 26, 2022).

[4] Government of India Ministry of Civil Aviation, Report of court of inquiry on mid-air collision between Saudi Arabian Boeing 747 and Kazakhstan IL-76, 1997. https://www.baaa-acro.com/sites/default/files/2020-12/UN-76435.pdf (accessed February 25, 2022).

[5] R. W. Mills, Incident reporting systems: Lessons from the Federal Aviation Administration's air traffic organization, 2013. https://www.businessofgovernment.org/sites/default/files/Incident%20Reporting%20Systems.pdf (accessed February 25, 2022). Could you please help us remove the link's underline and insert the corresponding hyperlink? Thanks!

[6] U.S. Department of Transportation, Key issues facing the federal aviation administration's controller workforce, 2008. https://www.oig.dot.gov/sites/default/files/WEB_FINAL_6-9-08_revised_July_2008.pdf (accessed February 5, 2022).

[7] S.M. Galster, J.A. Duley, A.J. Masalonis, R. Parasuraman, Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation, Int. J. Aviat. Psychol. 11 (2009) 71–93, https://doi.org/10.1207/S15327108IJAP1101_5.

[8] D. Dasari, G. Shou, L. Ding, ICA-Derived EEG correlates to mental fatigue, effort, and workload in a realistically simulated air traffic control task, Front. Neurosci. 11 (2017) 297, https://doi.org/10.3389/FNINS.2017.00297/BIBTEX.

[9] S.V. Ligda, M.J. Harris, C.S. Lieber, N.J. Cooke, Monitoring human performance in real-time for NAS safety prognostics, in: AIAA Aviation 2019 Forum, AIAA, 2019: pp. 1–8. https://doi.org/10.2514/6.2019-3411.

[10] T. Edwards, J. Homola, J. Mercer, L. Claudatos, Multifactor interactions and the air traffic controller: The interaction of situation awareness and workload in association with automation, IFAC-PapersOnLine. 49 (2016) 597–602, https://doi.org/10.1016/J.IFACOL.2016.10.623.

[11] J. Vogt, T. Hagemann, M. Kastner, The impact of workload on heart rate and blood pressure in en-route and tower air traffic control, J. Psychophysiol. 20 (2006) 297–314, https://doi.org/10.1027/0269-8803.20.4.297.

[12] F. Trapsilawati, M. Herliansyah, A. Nugraheni, M. Fatikasari, G. Tissamodie, EEG-based analysis of air traffic conflict: Investigating controllers' situation awareness, stress level and brain activity during conflict resolution, J. Navig. 73 (2020) 678–696, https://doi.org/10.1017/S0373463319000882.

[13] Q. Li, K.K.H. Ng, Z. Fan, X. Yuan, H. Liu, L. Bu, A human-centred approach based on functional near-infrared spectroscopy for adaptive decision-making in the air traffic control environment: a case study, Adv. Eng. Inf. 49 (2021), 101325, https://doi.org/10.1016/J.AEI.2021.101325.

[14] F. Li, C.H. Lee, C.H. Chen, L.P. Khoo, Hybrid data-driven vigilance model in traffic control center using eye-tracking data and context data, Adv. Eng. Inf. 42 (2019), 100940, https://doi.org/10.1016/J.AEI.2019.100940.

[15] H.J. Wee, S.W. Lye, J.P. Pinheiro, An integrated highly synchronous, high resolution, real time eye tracking system for dynamic flight movement, Adv. Eng. Inf. 41 (2019), 100919, https://doi.org/10.1016/J.AEI.2019.100919.

[16] M.N. Rastgoo, B. Nakisa, F. Maire, A. Rakotonirainy, V. Chandran, Automatic driver stress level classification using multimodal deep learning, Expert Syst. Appl. 138 (2019), 112793, https://doi.org/10.1016/J.ESWA.2019.07.010.

[17] Q. Ji, Z. Zhu, P. Lan, Real-time nonintrusive monitoring and prediction of driver fatigue, IEEE Trans. Veh. Technol. 53 (2004) 1052–1068, https://doi.org/10.1109/TVT.2004.830974.

[18] C.S. Lieber, M. Demir, N.J. Cooke, S. Ligda, Deviations in closed loop communications between air traffic controllers and pilots as a predictor of loss of separation, in: AIAA Aviation 2021 Forum, AIAA, 2021. https://doi.org/10.2514/6.2021-2320.

[19] Z. Sun, P. Tang, Automatic communication error detection using speech recognition and linguistic analysis for proactive control of loss of separation, Transp. Res. Rec. 2675 (2021) 1–12, https://doi.org/10.1177/0361198120983004.

[20] Y. Wang, P. Tang, Y. Shi, Y. Liu, N.J. Cooke, Predicting terminal mid-air collisions through simulator experiments of air traffic control, in: 2020 Winter Simulation Conference (WSC), IEEE, 2020: pp. 2536–2548. https://doi.org/10.1109/WSC48552.2020.9384047.

[21] J. Djokic, B. Lorenz, H. Fricke, Air traffic control complexity as workload driver, Transportation Research Part C: Emerging Technologies. 18 (2010) 930–936, https://doi.org/10.1016/j.trc.2010.03.005.

[22] Federal Aviation Administration, The air traffic controller workforce plan 2021-2030, 2020. https://www.faa.gov/air_traffic/publications/controller_staffing/media/2021-AFN_010-CWP2021.pdf (accessed June 3, 2022).

[23] S. Loft, P. Sanderson, A. Neal, M. Mooij, Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications, Hum. Factors 49 (2007) 376–399, https://doi.org/10.1518/001872007X197017.

[24] U. Metzger, R. Parasuraman, The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring, Hum. Factors 43 (2001) 519–528, https://doi.org/10.1518/001872001775870421.

[25] G.C. Fraccone, V. Volovoi, A.E. Colón, M. Blake, Novel air traffic procedures: Investigation of off-nominal scenarios and potential hazards, J. Aircr. 48 (2012) 127–140, https://doi.org/10.2514/1.C031003.

[26] Z. Kang, E.J. Bass, D.W. Lee, Air traffic controllers' visual scanning, aircraft selection, and comparison strategies in support of conflict detection, Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 58 (2014) 77–81. https://doi.org/10.1177/1541931214581017.

[27] K. Lee, E. Ferons, A. Pritchett, Describing airspace complexity: Airspace response to disturbances, J. Guid. Control Dynam. 32 (2012) 210–222, https://doi.org/10.2514/1.36308.

[28] A. Basu, J.S. Mitchell, G.K. Sabhnani, Geometric algorithms for optimal airspace design and air traffic controller workload balancing, Journal of Experimental Algorithmics 14 (2009) 2–3, https://doi.org/10.1145/1498698.1537598.

[29] C.D. Wickens, B.L. Hooey, B.F. Gore, A. Sebok, C.S. Koenicke, Identifying black swans in NextGen: Predicting human performance in off-nominal conditions, Hum. Factors 51 (2009) 638–651, https://doi.org/10.1177/0018720809349709.

[30] J. Orasanu, B. Parke, N. Kraft, Y.H. Tada, Alan Anderson, L. Barrett McDonnell, V. Dulchinos, Evaluating the effectiveness of schedule changes for air traffic service (ATS) providers: Controller alertness and fatigue monitoring study, 2012. https://www.faa.gov/data_research/research/media/nasa_controller_fatigue_assessment_report.pdf (accessed March 4, 2022).

[31] S.G. Hart, L.E. Staveland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, Adv. Psychol. 52 (1988) 139–183, https://doi.org/10.1016/S0166-4115(08)62386-9.

[32] A.J. Tattersall, P.S. Foord, An experimental evaluation of instantaneous self-assessment as a measure of workload, Ergonomics 39 (1996) 740–748, https://doi.org/10.1080/00140139608964495.

[33] K.A. Bernhardt, D. Poltavski, T. Petros, F.R. Ferraro, T. Jorgenson, C. Carlson, P. Drechsel, C. Iseminger, The effects of dynamic workload and experience on commercially available EEG cognitive state metrics in a high-fidelity air traffic control environment, Appl. Ergon. 77 (2019) 83–91, https://doi.org/10.1016/J.APERGO.2019.01.008.

[34] J. Xing, Z. Sun, P. Tang, A. Yilmaz, R. Laurids Boring, G.E. Gibson, Evaluating operator's real-time mental workload with eye movement analysis in nuclear power plants operations, in: in: Proceedings of the 2021 ASCE International Conference on Computing in Civil Engineering (I3CE2021), 2021, https://doi.org/10.1061/9780784483893.178.

[35] G. Borghini, G. di Flumeri, P. Aricò, N. Sciaraffa, S. Bonelli, M. Ragosta, P. Tomasello, F. Drogoul, U. Turhan, B. Acikel, A. Ozan, J.P. Imbert, G. Granger, R. Benhacene, F. Babiloni, A multimodal and signals fusion approach for assessing the impact of stressful events on air traffic controllers, Sci. Rep. 10 (2020) 1–18, https://doi.org/10.1038/s41598-020-65610-z.

[36] G. Borghini, A. Bandini, S. Orlandi, G. di Flumeri, P. Arico, N. Sciaraffa, V. Ronca, S. Bonelli, M. Ragosta, P. Tomasello, U. Turhan, B. Acikel, A. Ozan, J.P. Imbert, G. Granger, R. Benhacene, F. Drogoul, F. Babiloni, Stress assessment by combining neurophysiological signals and radio communications of air traffic controllers, in: 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE 2020 (2020) 851–854, https://doi.org/10.1109/EMBC44109.2020.9175958.

[37] S. Charbonnier, R.N. Roy, S. Bonnet, A. Campagne, EEG index for control operators' mental fatigue monitoring using interactions between brain regions, Expert Syst. Appl. 52 (2016) 91–98, https://doi.org/10.1016/J.ESWA.2016.01.013.

[38] J. Kuo, M.G. Lenné, R. Myers, A. Collard-Scruby, C. Jaeger, C. Birmingham, Real-time assessment of operator state in air traffic controllers using ocular metrics, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2017, pp. 257–261, https://doi.org/10.1177/1541931213601547.

[39] M.L. Chen, S.Y. Lu, I.F. Mao, Subjective symptoms and physiological measures of fatigue in air traffic controllers, Int. J. Ind. Ergon. 70 (2019) 1–8, https://doi.org/10.1016/J.ERGON.2018.12.004.

[40] J. Chen, Y. Liu, N. Cooke, P. Tang, Detecting anomalous behaviors of air traffic controllers in time series of facial expressions and head poses, in: AIAA Aviation 2019 Forum, AIAA, 2019, pp. 1–8, https://doi.org/10.2514/6.2019-3412.

[41] P. Liu, R. Xiong, P. Tang, Mining observation and cognitive behavior process patterns of bridge inspectors, in: in: Proceedings of the 2021 ASCE International Conference on Computing in Civil Engineering (I3CE2021), 2021, https://doi.org/10.1061/9780784483893.075.

[42] R. Xiong, P. Liu, P. Tang, Human reliability analysis and prediction for visual inspection in bridge maintenance, in, in: Proceedings of the 2021 ASCE International Conference on Computing in Civil Engineering (I3CE2021), 2021, https://doi.org/10.1061/9780784483893.032.

[43] Z. Liu, Y. Peng, W. Hu, Driver fatigue detection based on deeply-learned facial expression representation, J. Vis. Commun. Image Represent. 71 (2020), 102723, https://doi.org/10.1016/J.JVCIR.2019.102723.

[44] M. Ye, W. Zhang, P. Cao, K. Liu, Driver fatigue detection based on residual channel attention network and head pose estimation, Appl. Sci. 11 (2021) 9195, https://doi.org/10.3390/APP11199195.

[45] P. Liu, H.L. Chi, X. Li, J. Guo, Effects of dataset characteristics on the performance of fatigue detection for crane operators using hybrid deep neural networks, Autom. Constr. 132 (2021), 103901, https://doi.org/10.1016/J.AUTCON.2021.103901.

[46] I. Mehmood, H. Li, W. Umer, A. Arsalan, M. Saad Shakeel, S. Anwer, Validity of facial features' geometric measurements for real-time assessment of mental fatigue in construction equipment operators, Adv. Eng. Inf. 54 (2022), 101777, https://doi.org/10.1016/J.AEI.2022.101777.

[47] M. Rezaei, R. Klette, Look at the driver, look at the road: No distraction! No accident!, in: in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 129–136, https://doi.org/10.1109/CVPR.2014.24.

[48] L. Wang, Attention decrease detection based on video analysis in E-learning, in: Transactions on Edutainment XIV, Springer, 2018: pp. 166–179. https://doi.org/10.1007/978-3-662-56689-3_14.

[49] S. Hachisuka, Human and vehicle-driver drowsiness detection by facial expression, in: in: 2013 International Conference on Biometrics and Kansei Engineering, 2013, pp. 320–326, https://doi.org/10.1109/ICBAKE.2013.89.

[50] D. Liu, P. Sun, Y.Q. Xiao, Y. Yin, Drowsiness detection based on eyelid movement, in,, Second International Workshop on Education Technology and Computer Science, IEEE 2010 (2010) 49–52, https://doi.org/10.1109/ETCS.2010.292.

[51] A. Mittal, K. Kumar, S. Dhamija, M. Kaur, Head movement-based driver drowsiness detection: A review of state-of-art techniques, in: 2016 IEEE International Conference on Engineering and Technology (ICETECH), IEEE, 2016: pp. 903–908. https://doi.org/10.1109/ICETECH.2016.7569378.

[52] H. Gao, A. Yuce, J.P. Thiran, Detecting emotional stress from facial expressions for driving safety, in: in: 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 5961–5965, https://doi.org/10.1109/ICIP.2014.7026203.

[53] L. Zhang, S. Qin, Z. Yao, K. Zhang, J. Wu, Long-term academic stress enhances early processing of facial expressions, Int. J. Psychophysiol. 109 (2016) 138–146, https://doi.org/10.1016/J.IJPSYCHO.2016.08.010.

[54] C. Jyotsna, J. Amudha, Eye gaze as an indicator for stress level analysis in students, in: in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018, pp. 1588–1593, https://doi.org/10.1109/ICACCI.2018.8554715.

[55] G. Giannakakis, D. Manousos, V. Chaniotakis, M. Tsiknakis, Evaluation of head pose features for stress detection and classification, in: in: 2018 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)), 2018, pp. 406–409, https://doi.org/10.1109/BHI.2018.8333454.

[56] B. Zheng, X. Jiang, G. Tien, A. Meneghetti, O.N.M. Panton, M.S. Atkins, Workload assessment of surgeons: Correlation between NASA TLX and blinks, Surg. Endosc. 26 (2012) 2746–2750, https://doi.org/10.1007/S00464-012-2268-6/TABLES/2.

[57] H.J. Foy, P. Chapman, Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation, Appl. Ergon. 73 (2018) 90–99, https://doi.org/10.1016/J.APERGO.2018.06.006.

[58] H. Jebelli, S. Hwang, S. Lee, EEG-based workers' stress recognition at construction sites, Autom. Constr. 93 (2018) 315–324, https://doi.org/10.1016/j.autcon.2018.05.027.

[59] H. Liu, X. Mi, Y. Li, Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM, Energ. Conver. Manage. 159 (2018) 54–64, https://doi.org/10.1016/J.ENCONMAN.2018.01.010.

[60] J. Cai, Y. Zhang, L. Yang, H. Cai, S. Li, A context-augmented deep learning approach for worker trajectory prediction on unstructured and dynamic construction sites, Adv. Eng. Inf. 46 (2020), 101173, https://doi.org/10.1016/J.AEI.2020.101173.

[61] T. Jing, P. Zheng, L. Xia, T. Liu, Transformer-based hierarchical latent space VAE for interpretable remaining useful life prediction, Adv. Eng. Inf. 54 (2022), 101781, https://doi.org/10.1016/J.AEI.2022.101781.

[62] M. Prandini, L. Piroddi, S. Puechmorel, S.L. Brázdilová, Toward air traffic complexity assessment in new generation air traffic management systems, IEEE Trans. Intell. Transp. Syst. 12 (2011) 809–818, https://doi.org/10.1109/TITS.2011.2113175.

[63] M. Patel, S.K.L. Lal, D. Kavanagh, P. Rossiter, Applying neural network analysis on heart rate variability data to assess driver fatigue, Expert Syst. Appl. 38 (2011) 7235–7242, https://doi.org/10.1016/J.ESWA.2010.12.028.

[64] T.Z. Strybel, K.P.L. Vu, D.L. Chiappe, C.A. Morgan, G. Morales, V. Battiste, Effects of NextGen concepts of operation for separation assurance and interval management on air traffic controller situation awareness, workload, and performance, International Journal of Aviation Psychology. 26 (2016) 1–14, https://doi.org/10.1080/10508414.2016.1235363.

[65] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.P. Morency, OpenFace 2.0: Facial behavior analysis toolkit, in: in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 2018, pp. 59–66, https://doi.org/10.1109/FG.2018.00019.

[66] T. Baltrušaitis, P. Robinson, L.P. Morency, Constrained local neural fields for robust facial landmark detection in the wild, in: in: 2013 IEEE International Conference on Computer Vision Workshops, 2013, pp. 354–361, https://doi.org/10.1109/ICCVW.2013.54.

[67] T. Soukupová, J. Cech, Real-time eye blink detection using facial landmarks, in: Proceedings of the 21st Computer Vision Winter Workshop, Slovenian Pattern Recognition Society, 2016: pp. 1–8. https://dspace.cvut.cz/bitstream/handle/10467/64839/F3-DP-2016-Soukupova-Tereza-SOUKUPOVA_DP_2016.pdf (accessed November 21, 2021).

[68] O. Arriaga, M. Valdenegro-Toro, P.G. Plöger, Real-time convolutional neural networks for emotion and gender classification, ArXiv Preprint. (2017). https://arxiv.org/abs/1710.07557v1 (accessed November 21, 2021).

[69] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[70] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient convolutional neural networks for mobile vision applications, ArXiv Preprint. (2017). https://arxiv.org/abs/1704.04861v1 (accessed November 21, 2021).

[71] S.T. Shorrock, B. Kirwan, Development and application of a human error identification tool for air traffic control, Appl. Ergon. 33 (2002) 319–336, https://doi.org/10.1016/S0003-6870(02)00010-8.

[72] International Civil Aviation Organization, Procedures for air navigation services — Air traffic management (Doc 4444), 2016. https://ops.group/blog/wp-content/uploads/2017/03/ICAO-Doc4444-Pans-Atm-16thEdition-2016-OPSGROUP.pdf (accessed December 22, 2021).

[73] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/NECO.1997.9.8.1735.

[74] MetaCraft, MetaCraft: ATC training software & solutions, https://www.metacraft.com/ (accessed November 19, 2021).

[75] Advanced Brain Monitoring (ABM), B-Alert X10, https://www.advancedbrainmonitoring.com/products/b-alert-x10 (accessed December 25, 2022).

[76] J. Li, H. Li, H. Wang, W. Umer, H. Fu, X. Xing, Evaluating the impact of mental fatigue on construction equipment operators' ability to detect hazards using wearable eye-tracking technology, Autom. Constr. 105 (2019), 102835, https://doi.org/10.1016/j.autcon.2019.102835.

[77] S. Yan, C.C. Tran, Y. Wei, J.L. Habiyaremye, Driver's mental workload prediction model based on physiological indices, Int. J. Occup. Saf. Ergon. 25 (2019) 476–484, https://doi.org/10.1080/10803548.2017.1368951.

[78] Y. Hochberg, A sharper Bonferroni procedure for multiple tests of significance, Biometrika 75 (1988) 800–802, https://doi.org/10.1093/BIOMET/75.4.800.

[79] J. Hauke, T. Kossowski, Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data, Quaestiones Geographicae. 30 (2011) 87–93, https://doi.org/10.2478/V10117-011-0021-1.

[80] J.D. Rodríguez, A. Pérez, J.A. Lozano, Sensitivity analysis of k-fold cross validation in prediction error estimation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 569–575, https://doi.org/10.1109/TPAMI.2009.187.

[81] W.R. Pearson, An introduction to sequence similarity ("homology") searching, Curr. Protoc. Bioinformatics 42 (2013), https://doi.org/10.1002/0471250953.BI0301S42.

[82] J. Paxion, E. Galy, C. Berthelon, Mental workload and driving, Front. Psychol. 5 (2014) 1344, https://doi.org/10.3389/FPSYG.2014.01344/BIBTEX.

[83] Y.J. Yao, Y.M. Chang, X.P. Xie, X.S. Cao, X.Q. Sun, Y.H. Wu, Heart rate and respiration responses to real traffic pattern flight, Applied Psychophysiology, Biofeedback 33 (2008) 203–209, https://doi.org/10.1007/S10484-008-9066-X/TABLES/1.