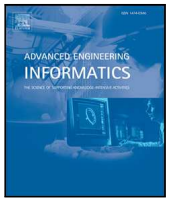
Advanced Engineering Informatics 57 (2023) 102113

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Full length article

Air traffic controller workload level prediction using conformalized dynamical graph learning

Yutian Pang [a], Jueming Hu [a], Christopher S. Lieber [b], Nancy J. Cooke [b], Yongming Liu [a,*]

[a] Mechanical and Aerospace Engineering, Arizona State University, Tempe, 85287, AZ, USA
[b] Human Systems Engineering, Arizona State University - Polytechnic, Mesa, 85212, AZ, USA

ARTICLE INFO

Keywords:
Air traffic management
Aviation human factors
Controller workload
Graph neural network

ABSTRACT

Air traffic control (ATC) is a safety-critical service system that demands constant attention from ground air traffic controllers (ATCos) to maintain daily aviation operations. The workload of the ATCos can have negative effects on operational safety and airspace usage. To avoid overloading and ensure an acceptable workload level for the ATCos, it is important to predict the ATCos' workload accurately for mitigation actions. In this paper, we first perform a review of research on ATCo workload, mostly from the air traffic perspective. Then, we briefly introduce the setup of the human-in-the-loop (HITL) simulations with retired ATCos, where the air traffic data and workload labels are obtained. The simulations are conducted under three Phoenix approach scenarios while the human ATCos are requested to self-evaluate their workload ratings (i.e., low-1 to high-7). Preliminary data analysis is conducted. Next, we propose a graph-based deep-learning framework with conformal prediction to identify the ATCo workload levels. The number of aircraft under the controller's control varies both spatially and temporally, resulting in dynamically evolving graphs. The experiment results suggest that (a) besides the traffic density feature, the traffic conflict feature contributes to the workload prediction capabilities (i.e., minimum horizontal/vertical separation distance); (b) directly learning from the spatiotemporal graph layout of airspace with graph neural network can achieve higher prediction accuracy, compare to hand-crafted traffic complexity features; (c) conformal prediction is a valuable tool to further boost model prediction accuracy, resulting a range of predicted workload labels. The code used is available at Link.

1. Introduction

The rapid advancement of intelligent systems substantially reduces the operational effort from the individual user level but escalates the system-level complexity of real-time decision-making and corporate planning, arises from the dynamically changing environments, time restrictions, and tactical constraints [1–4]. Workload assessment and prediction of operating such complex systems have long been regarded as critical research objects [5–9]. Workload overhead can occur when the demands exceed the human operator's capacity and can lead to efficiency drop and operational safety concerns. Within the aviation domain, effective workload management of air traffic controllers (ATCos) is of utmost importance to maintain safety and rely on accurate ATCo workload predictions.

Air traffic control (ATC) is a crucial part of aviation safety, ensuring that aircraft are safely guided through the airspace and landed or taken off from airports. ATCos are responsible for managing the flow of aircraft, communicating with pilots, and making critical decisions in real-time to ensure the safety of all involved. As air traffic continues to increase [10], it puts more pressure on ATCos, who already have highly demanding and stressful daily routines [11–13]. Quantifying the effort made to meet these task requirements lead to the concept of workload as an air traffic controller, which denotes the subjective qualitative measure of perception demand placed by the current air traffic situation [4,7,14]. Moreover, proper workload management and scheduling are vital to ensure ATCos can perform their duties effectively and faultlessly without being overwhelmed. Human performance is a crucial factor in ensuring the safe operations of the National Airspace System (NAS). In the past, human operators have been identified as significant contributors to accidents involving air carrier operations governed by the 14 Code of Federal Regulations (CFR) Part 121, which covers commercial airliners frequently used by the public [15]. For instance, approximately 80% of the 446 air carrier accidents that occurred between 1997 and 2006 were attributed to personnel-related factors, while environmental factors were cited in approximately 40% of the accidents and aircraft-related factors in 20% [16]. Workload prediction can help in several ways, such as ensuring that enough ATCos are

* Corresponding author.
E-mail address: Yongming.liu@asu.edu (Y. Liu).

https://doi.org/10.1016/j.aei.2023.102113
Received 16 April 2023; Received in revised form 18 June 2023; Accepted 18 July 2023
Available online 5 August 2023

available to manage the traffic, preventing fatigue and burnout, optimizing shift schedules, and improving overall efficiency. By accurately predicting the workload, air traffic organizations can ensure that they have the necessary resources and personnel to maintain a safe and efficient air traffic system [17–19]. All of these objectives are based on reliable ATCo workload-level modeling and predictions. Moreover, artificial intelligence (AI)-enabled human factor studies in aviation have been identified as one of the core elements of AI taxonomy by related authorities [20,21], where ATCo workload management is a key dedicated objective.

A tremendous amount of research has been done to understand the impact factors and demand patterns that drive the workload of a controller, such that a better workload prediction performance can be discovered. Two types of factors are studied extensively in the literature, (a) physiological and behavioral features including ATCo mental stress, fatigue level, communication difficulties, and situation awareness [22–25]; (b) objective factors such as traffic and airspace complexity measures (i.e., operational errors (OE)), abnormal events, level of automation, and weather situations [26–32]. In order to collect features for workload prediction, researchers have proposed to collect human-subject data (i.e., eye movement, communications, heartbeat rates, and Electroencephalography (EEG) signals) [33–35], in an intrusive and non-intrusive sense. On the other hand, traffic-related features can be directly obtained from computer flight recordings and operational recordings. However, some specific traffic features need *post-hoc* processing (i.e., loss of separations (LoS), OEs). Specifically, in this work, we are interested in traffic-related objective features since it is very unlikely to collect real-time biological features (i.e., EEG/ECG signals or heartbeat rates) in the near future due to privacy concerns and regulatory requirements. Moreover, we discover that the existing model on workload prediction is mostly using handcrafted features, even if a graph data structure and simple neural networks (i.e., minimum spanning trees) have been proposed [27]. To the best of the authors' knowledge, there is no investigation on utilizing advanced data-driven learning techniques (i.e., graph neural networks (GNNs)) for workload prediction possibilities that directly leverage the spatiotemporal relationships contained in the traffic data and airspace layout.

In this work, we investigate the possibility of using the graph neural network to predict ATCo workload levels, with an additional post-processing technique, namely conformal prediction, to boost the accuracy with a set of prediction labels. The data is collected by conducting experiments with retired ATC participants who have experience at FAA Radar Approach Control (TRACON) facilities, under three different scenarios, (a) baseline conditions; (b) high workload nominal conditions; (c) high workload off-nominal conditions. The major difference among scenarios are the peak traffic densities and the presence of off-nominal events (i.e., runway switch, communication errors, etc.). The simulation scenario is limited to a few Phoenix approach procedures for a duration of 25 min for each scenario. A detailed description of the experimental setup is in Section 3. Specifically, the ATCo workload we investigated is the executive (R-side) controllers' workload [17]. Predicting controller workload levels can be viewed as a pattern recognition problem [28] and thus is suitable for data-driven learning algorithms. In this work, the problem of predicting workload based on the spatiotemporal layout of airspace is viewed as a time-series dynamically evolving graph classification task. Being time-series classification, we propose to input multiple historical timestamp graphs into the model for the prediction of workload level at the next timestamp. Also, the spatiotemporal layout of the graph structure varies at each timestamp (i.e., number of nodes, graph edge connections), resulting in a dynamical graph classification problem.

Our contributions are summarized as,

- This paper investigates the possibility of predicting executive controller workload during approach scenarios directly from the recorded air traffic data with graph neural networks and discovers that traffic conflict is a nontrivial contributor to improving workload prediction capabilities.
- We propose to formulate the ATCo workload prediction task into a dynamical time-series graph classification problem and show that the Evolving Graph Convolutional Network (EvolveGCN) can achieve a higher prediction accuracy than both statistical (i.e., regression, handcrafted features) and classical learning methods (i.e., MLP, GCN). We show that graph neural networks have great potential for predicting controller workload with varying spatiotemporal airspace layouts.
- A moving window approach is proposed to build the correct input–output matching from the collected sparse workload data. The moving window size represents the temporal length of the historical information used in workload prediction. The selection of parameters can be alternated to fit into the operational need. The data structure formulation transfer complex structured traffic features into a lucid format for research and development purposes.
- To further improve the classification accuracy of the experimental data. We explore conformal prediction to expand the prediction as set predictions. We show that conformal prediction has better ground truth label coverage by giving multiple possible predictions as indicators of model uncertainties. We suggest that conformal prediction is a valuable machine learning *post-hoc* processing tool to boost performance further as well as indicate prediction uncertainties.

The rest of the paper is organized as follows. First, Section 2 reviews related studies on air traffic controller workload prediction. We first introduce the impact factors of ATCo workload in Section 2.1, then list the current practices in predicting workload Section 2.2. In Section 3, we introduce the detailed workflow of human-in-the-loop simulations to collect the traffic data and ground truth ATC workload labels, along with data analysis of the collected data. Section 4 describes the flowchart of the proposed machine learning framework, from experiment data handling to innovative modeling. The prediction performance and evaluation of the conformal prediction set are discussed in Section 5. Section 6 concludes this paper by giving limitations of this study and provides future insights.

2. Related works

Due to the surging number of daily aviation operations, the aviation industry is in urgent need of advanced decision support tools that can accommodate the rapid annual air traffic growth. Numerous studies have investigated ATCo workload. It is an aviation researchers' consensus that understanding the impact factors that drive mental workload can help improve airspace capacity, thus reducing aviation safety concerns [4,36,37]. With meaningful impact factors collected or modeled, predictive modeling is critical to building an accurate workload prediction algorithm. In this section, we discuss the related works from two aspects, (1) understand the impact factors that drive the mental workload in Section 2.1; (2) discuss the current practice in predicting ATCo workload from open literature with a focus on predicting workload from traffic factors Section 2.2.

2.1. Task demands and impact factors to ATCo workload

In air traffic control, task demand refers to the level of mental and physical effort required for ATCos to complete their duties effectively. High task demand leads to increased workload, which negatively impacts aviation safety. Therefore, understanding the level of task demand and appropriately managing workload is essential for ensuring that ATCos can perform their duties effectively and safely. Correctly modeling task demand is viewed as the prerequisite for workload prediction for a long history. It is noteworthy to mention that ATCos workload is not a simple function of task demands; the ATC strategy the controller adapted to meet the increased task demands also provides a feedback

loop to ATCo workload [4]. We discuss each of these aforementioned grouped impact factors separately.

Air traffic factors refers to both the aircraft count under the ATCo control and their spatiotemporal relationships. The number of aircraft under control is viewed as the most important factor that drives ATCo's mental workload [38,39]. A high aircraft count leads to higher communication frequency and a higher possibility of safety events, resulting in a higher mental and physical workload. Traffic density is typically measured in aircraft per unit of airspace, such as aircraft in unit time and unit airspace sector area. Measurements of traffic density have been developed based on the averaged vertical/horizontal separation distances [28], as they directly infer loss of separations. Other research investigates the necessity to consider flight interactions and flight characteristics, which includes the changes/variability in heading, speed, or altitude, the pattern of how air traffic flows merges and separates into a set of air traffic complexity metrics [7,28,39–41]. Their regression analysis shows subjective workload depends on both aircraft count and other air traffic complexity measures. Additionally, some other studies also suggest that a lower aircraft count also can lead to task overload if these aircraft are interacting in a complex fashion [42,43].

Airspace complexity factors include the number of routes, altitudes, and restrictions, which can also impact the workload of ATCos, as they need to monitor and manage multiple variables simultaneously. Airspace-related factors are another key contributor to ATCo mental workload [44]. Larger airspace size indicates a higher aircraft count and higher metrics on traffic complexity, while small airspace size reduces conflict resolution options and higher traffic evolving rates. It is noteworthy to mention another work considering both the traffic factors and airspace structure complexity and proposes Structural Complexity Metric (SCM), which incorporates a measure of the organization, hierarchy, and interdependence into the complexity calculation [45]. Furthermore, this paper suggests using well-defined ingress and egress points in the airspace to distinguish normal and abnormal flights based on real-time monitoring.

Operational Constraints are another major contributor that drives the ATCo workload. Operational constraints refer to the temporal variability within the operational conditions of the airspace, as well as the conditions of related technology and equipment. Several factors are viewed as operational constraints; (1) pilot-controller communications are critical for maintaining safe aviation operations. Malfunctions of communication devices can disrupt air traffic control operations. This is known as loss of ratio communication (NORDO) (2) convective weather conditions, such as thunderstorms or heavy fog. These types of objective factors can affect air traffic control operations by reducing visibility and creating unsafe flying conditions. (3) subjective airspace restriction is another type of operation constraint. The restrictions come from multiple sources, i.e., aircraft holding, no-fly zones, or special-use airspace [4]. In addition, certain other off-nominal events are considered operational constraints, i.e, runway switch, and minimum fuel reported [46,47].

Cognitive states directly contribute to the cognitive task demands of ATCo. To measure cognitive states, the researchers propose to measure the physiological states of the air traffic controller, including brain activities, eye movements, and heartbeat rates. These states can be quantitatively measured by sensors signals such as electrocardiography (ECG) signals, electroencephalography (EEG) signals, galvanic skin response (GSR), blood pressure (BP), and certain biochemical analysis [48–50]. However, using intrusive physiological state measurements is disruptive to controllers' normal working conditions, as it creates additional mental stress and discomfort in maintaining ATC operations. Alternatively, computer vision (CV) based non-intrusively physiological state measurements are proposed to collect distractions, drowsiness, head poses, eye movements, and fatigue levels [19,51,52]. However, these types of measurements can lead to information security and privacy concerns [53,54].

2.2. Workload prediction algorithms

The dynamic density model builds a regression model to find the linear relationships between traffic complexity factors and ATCo workload. The Dynamic Density metric uses a combination of traffic density and complexity measures to estimate task demands in real-time, with the goal of providing a more accurate and responsive measure of controller workload from task demands [55]. However, dynamic density metrics fail to consider human cognitive capacities, which are the primary source of ATCo workload sources in the real world. In [39], the traffic complexity, as well as the airspace complexity of different sections, are considered. The results show that the airspace factor can actually contribute to workload prediction in a multi-sector study. Similarly, in [23], the authors find that the ATCo workload is proportional to the number of aircraft controlled by the enroute sector. They conduct HITL experiments and found a linear relationship between aircraft count and workload ratings.

However, it is still difficult to identify the most contributing impact factors to workload prediction from regression analysis. The first reason is the multi-collinearity within these factors. For instance, the number of conflicts depends on the speed, altitude, and heading variabilities. The complexity of traffic situations, such as traffic density and potential conflicts, also mediate the causal connection between traffic count and workload. The inter-relationship of these factors makes it challenging to determine the relative importance of each predictor in a regression equation. The second query is the debate on the linear relationship between these factors and workload ratings. For instance, the ATCo can alternate control strategies during a certain period to meet the increased task demands. Additionally, the online processing or post-processing of these factors only considers the current situations, and there is no inference on ATCo's intent and air traffic intent information. Trajectory prediction from flight plans helps with estimating the workload of ATCos, where prediction and reduction of trajectory uncertainties can help alleviate ATCo workload [31,56].

Machine learning algorithms have been adopted for workload prediction. In [57], tree-based models, and support vector machines are included for workload measurements. The authors consider both traffic complexity and operational constraints as features and show a high F1 score of over 0.9. However, these types of works are not dynamically considering the *traffic pattern* of the airspace, but still formulate the data as a tabular format, which fails to address the aforementioned concerns [4]. On the other hand, [9] proposes to use a 3-layer simple neural network to forecast ATCo workload. However, the ATCo workload in this work is assessed from voice communication data, which fails to model the task demand factors mentioned earlier. Impressively, direct prediction from the spatiotemporal air traffic layout is actually proposed decades ago. In [26,28], the authors propose to model air traffic at each timestamp into graph-structured data and calculate the second-order statistics of time-series of graphs as extracted air traffic complexity measurements. Then a simple neural network is adopted to do classification from these features.

In summary, understanding the factors that drive ATCo workload has been a challenging yet unresolved open question for decades. Instead of investigating the linear relationship between impact factors and workload, one should look into the dynamic properties of these factors and workload. This leads to our study on workload prediction — we model the spatiotemporal airspace layout into time-series graph structures and propose to use a time-series learning algorithm to predict workload levels, in consideration of historical dynamic variabilities contained within the air traffic data.

3. Human-in-the-loop (HITL) simulations

In this section, we provide an overview and the detailed simulation setup of the Human-In-The-Loop (HITL) experiment as in Fig. 1. The scope and objectives will be discussed. Then, data analysis on the collected data is presented in Section 3.3.

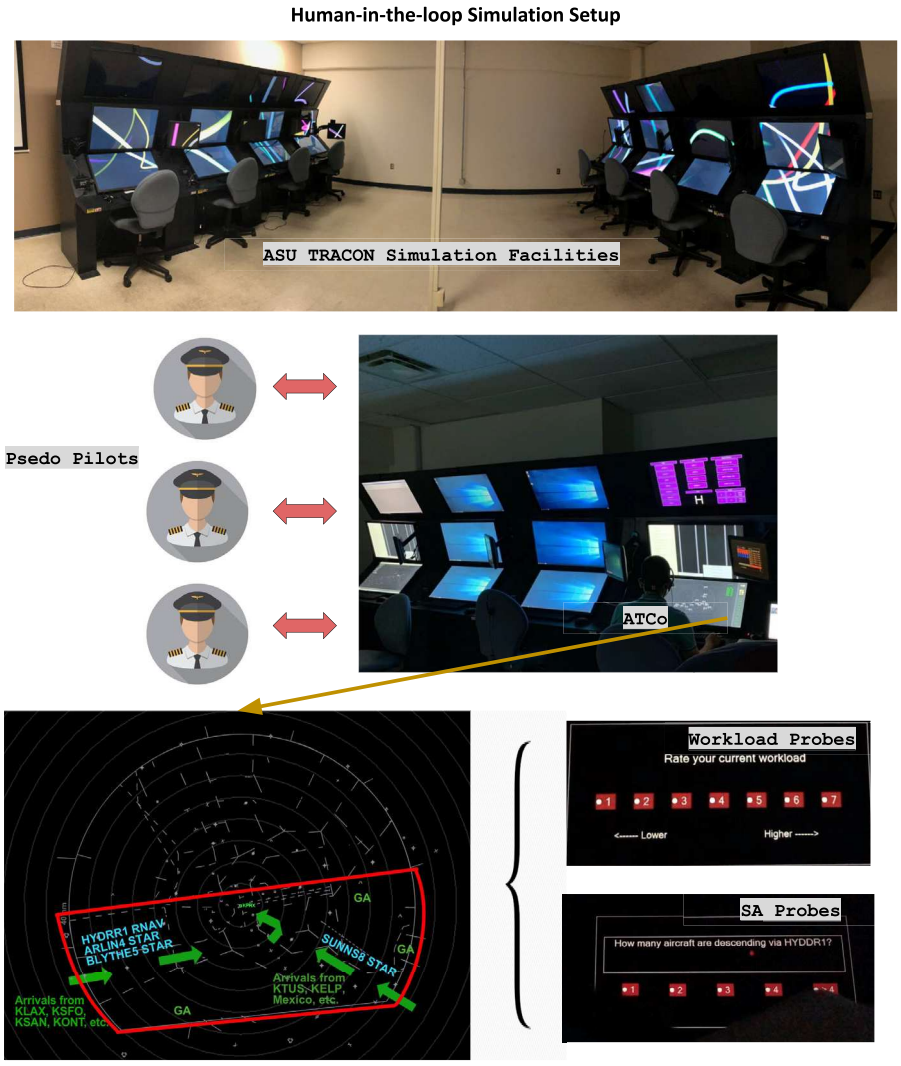


Fig. 1. Human-In-The-Loop (HITL) experiment setup. The ASU TRACON Simulation facility is equipped with eight simulators. During the HITL experiments, three pseudo pilots act as pilots and interact with ATCo. We show a simple demonstration of the graphical user interface, where the primary focus of the simulation is on the KPHX arrivals from two directions, with flight procedures including 1 RNAV (HYDRR1) and 3 STARs (ARLIN4, BLYTHE5, SUBSS8). During the experiment, the ATCo is asked to respond to the questions shown on the pop-up window. The window acts as either a workload probe or a situational awareness probe, showing every 3 min.

3.1. Simulation overview

The HITL simulation is the first human factor study of our aviation big data project, which aims at addressing the safety needs and technology solutions for future NAS [58]. The backbone of the project lies in information fusion and uncertainty management for real-time system-wide safety assurance, where human factors like ATCo workload play a key role. As mentioned, accurately predicting the ATCos workload can improve operational efficiency and reduce safety concerns such that aviation authorities can ensure a reasonable resource allocation and workload management.

The primary objective of the HITL experiment is to investigate the correlation between communication patterns (such as content, volume, and flow patterns) and both controller workload and human performance. Fig. 1 gives the overview setup of the HITL simulation process. The simulation contains two arrival flows including four Phoenix inbound Procedures. The top panel of Fig. 1 displays the layout of the ASU TRACON Simulation facility. The ASU TRACON Simulation facility boasts a total of eight advanced simulators, designed to provide an immersive training experience. Within the context of HITL (Human-in-the-Loop) experiments, the simulation involves the participation of three pseudo-pilots who assume the role of actual pilots and engage in interactive communication with Air Traffic Controllers (ATCo).

To illustrate the capabilities of the simulation, we present a captivating demonstration of the graphical user interface. The simulation places particular emphasis on the KPHX (Phoenix Sky Harbor International Airport) arrivals from two distinct directions. The first arrival flow represents flights from the west coast (KLAX, KSFO, KSAN, KONT, etc.), with procedures HYDRR1, ARLIN4, and BLYTHE5. The second flow stands for arrival flights from the southeast, including KTUS, KELP, and Mexico. SUNSS8 arrival procedure is used here. These procedures challenge both pilots and ATCos to execute precise maneuvers and coordinate their actions effectively.

Throughout the experiment, the ATCo is presented with a series of thought-provoking questions displayed in a pop-up window. This window either acts as both a workload probe to test the ATCo's ability to manage multiple tasks or a situational awareness probe, evaluating their understanding of the ongoing situation. The questions are presented at regular intervals of three minutes, ensuring a continuous evaluation of the ATCo's performance and cognitive response. Each retired ATCo participant engaged in a within-subjects (3 simulation trials) study design. Each trial spanned 25 min and varies in workload level

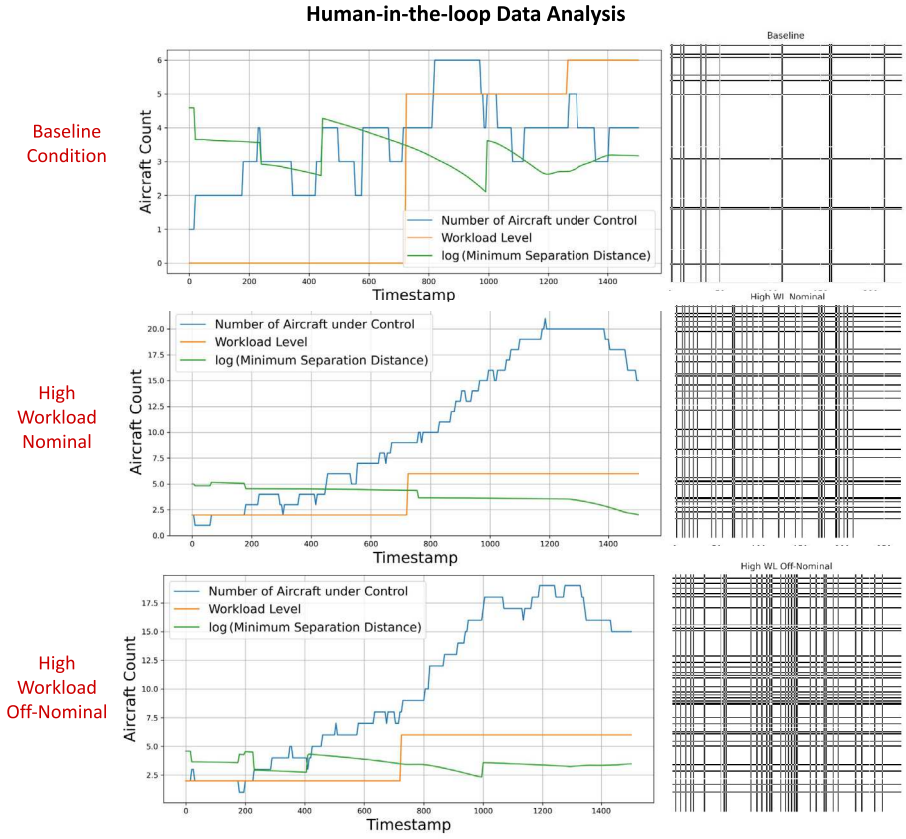


Fig. 2. The visualization of collected data samples from the HITL experiments under three scenarios. Each ATCo was involved in three experiments under normal, high workload nominal. high workload off-nominal working scenarios, with a time duration of 25 min each. Traffic-related features (e.g., aircraft numbers, minimum separation distance) are collected every 5 s. Workload levels are interpolated to match the corresponding traffic-related features at each timestamp. Additionally, we obtain the recurrence plot (RP) from communication transcripts as indicators of system tendency in Fig. 3.

Table 1
A short view of the communication transcripts post-processed from radio recordings in HITL simulation. Three cut-off sections are listed here, which correspond to off-nominal events, (1) turbulence reported; (2) no radio communications; (3) landing runway switch. Indicator of communication deviations is also shown in the right-most column. A mark of "0" if the ATCo communication followed a pilot's communication or vice versa, and "1" will be noted if the communications are not effective and are immediately followed between the ATCo and the pilot.

Speaker	Start time	Transcriptions	End time	Deviation indicators
		... ...		
Speed bird 281	05:26.9	approach speed bird two eighty-one we are experiencing motor turbulence at one three thousand	05:30.5	1
PHX approach (Speed bird 281)	05:34.9	speed bird two eighty-one heavy, roger	05:37.7	1
		... ...		
PHX approach (Cumpacity 250)	11:32.8	cumpacity two fifty descending maintain four thousand one hundred	11:35.4	0
PHX approach (Cumpacity 250)	11:38.8	cumpacity two fifty four thousand one hundred	11:40.5	1
Cumpacity 250	11:41.8	cumpacity two fifty dropping down to four one hundred	11:44.6	1
		... ...		
Local south	14:29.1	court this is local south we are switching to runway seven left and right effect immediately	14:35.9	1
PHX approach (Local south)	14:38.8	Ok ... move to runway seven	14:40.7	1
PHX approach (Ascer 4527)	14:49.8	shuttle forty-five twenty-seven expect dail is runway seven left turn left in two seven zero maintain five thousand	14:55.1	1
		... ...		

by manipulating two variables, namely traffic density, and occurrence of off-nominal events.

Baseline: Baseline trials contain up to 6 aircraft in the airspace at a given timestamp. There are no off-nominal events in baseline trials. Typically, a moderate workload is expected.

High Workload Nominal: High workload nominal trials can have up to 21 aircraft showing up in the current simulation environment. Again, there are no off-nominal events.

High Workload Off-Nominal: In addition to the experimental setup in high workload nominal trials, high workload off-nominal trials incorporate four off-nominal events during the 25 min duration. We list the name of these off-nominal events here,

- Turbulence: Moderate turbulence is simulated in several arrival flows. In Table 1, speed bird 281 reported experiencing turbulence at a certain altitude, starting from 05:26.9 of simulation time.
- No radio (NORDO): The pilot has no radio communication with the approach ATCo, which can happen during a radio failure. In Table 1, the KPHX approach controller repeated the order when the first order at 11:32.8 was not confirmed.
- Runway switch: In this simulation, the landing runway switch from KPHX 25L to 07R. The order is given by the local tower, as in Table 1 14:29.1.
- Minimum fuel: At the end of each simulation trial, the aircraft encountered fuel issues.

These corresponding timestamps t_1, t_2, t_3, and t_4 to apply these off-nominal events is indicated in Fig. 3, respectively.

3.2. Simulation setup

HITL was conducted in eight air traffic management system Metacraft facilities located at the Arizona State University TRACON Simulation Lab, which can be operated as either ATC terminal radar positions or pseudo-pilot stations. The human controllers were retired ATCos who have experience with civilian TRACON facilities within the past 15 years but do not persist possess experience with Phoenix TRACON [11]. Six retired ATCos were involved in this study. There were also three researchers who act as pseudo-pilots to fly along the assigned arrival routes during each simulation scenario. Metacraft is the name of the TRACON radar simulation computer cluster system. It provides ATC functions to maneuver the aircraft in a simulation environment, including altitude, speed, and heading. Metacraft collects and maintains data logs such as spatiotemporal tracks, LoS events, and distance measures.

During each 25 min experiment trial, the pop-up window showed a questionnaire probe every 3 min, asking either a question on workload rating or situational awareness questions. Specifically, the workload rating questions were shown three times at exact 3 min, 12 min, and 21 min timestamps. Fig. 2 visualizes the collected data for one participant of three scenarios. Features include minimum separations (traffic conflict), number of aircraft (traffic density), and workload ratings are reported. A recurrence plot indicating communication tendency is also provided, visually representing the communication efforts between the tower controller and the pseudo-pilots.

The workload rating probe was designed based on the subjective workload assessment technique (SWAT) [59]. The SWAT method is a situation-present assessment method that is commonly used in human factors research. It involves participants rating their perceived workload using various rating scales, such as the NASA task load index (TLX) scale [60,61]. The SWAT method also includes measures of mental effort and task difficulty to provide a more comprehensive assessment of workload. The workload probe employs a two-step process for administering questions related to situation awareness or workload. Participants first press a ready button, followed by selecting a response. The timing of both actions is recorded, following the methods used in the aforementioned studies. Another important measure of SWAT was the behavioral measure of workload, the time to respond to the ready button.

The controller workload is self-evaluated by the workload question pop-up window, and the human performance is indicated by the count of separation violations. To facilitate this investigation, we have gathered preliminary data on three types of metrics: (1) aeronautical separation violations, which are viewed as traffic conflicts existed in the airspace and an indicator of ATCo performances; (2) real-time workload ratings, taken at three different points in each 25-minute scenario; and (3) audio recordings of controller-pilot transmissions during the workload ratings. The description of selected features is in Table 2. These initial settings will serve as a foundation for further analyses using additional measures, such as facial recognition, heart rate variability, situation awareness probes, and operational efficiency. Ultimately, this research provides a solution for the development of a real-time controller workload level prediction system.

In this paper, we obtain the flight traffic recordings and the real-time workload ratings from real-world human-in-the-loop simulations. The subjective workload rating is collected from the question pop-up windows showing at 3 min, 12 min, and 21 min for each trial. The originally collected data and adjusted workload rating score has been discussed in the literature [62]. By doing this, we obtain realistic human workload levels, or the ground truth, from participants' honest ratings of their mental status. This is the most reasonable data source for obtaining features and labels for building real-world machine-learning pipelines.

3.3. Empirical data analysis

Communication transcription analysis is performed based on post-processed radio recordings. There are three major components in this part, (a) use a speech recognition tool or manual transcription tool to translate voice to text; (b) identify the named entity of each communication transcript (i.e., controllers or pilots); (c) perform either statistical analysis or keyword extraction [63–65]. As mentioned, the deviation indicator represents the deviation in communications, also known as closed loop communication deviation (CLCD) [13]. CLCD is based on an established coding scheme derived from the expected exchange of closed loop communication (CLC). Deviations occur when consecutive pilot communications or consecutive air traffic controller communications take place. CLCs were coded using a binary system to detect CLCD based on communication patterns between pilots and ATC. An expected CLC pattern involves alternating communications between pilots and ATC. CLCD is identified when a pilot's communication follows another pilot's or when consecutive ATC communications occur.

Fig. 3 shows the closed loop communication deviation analysis's recurrence plot (RP). RP was originally proposed to visualize the complexity of dynamical systems [67,68], where a detailed mathematical formulation can be found. In this work, we define the phase vector as the communication deviations and build the recurrence matrix $\mathbf{R}$ as,

$$\mathbf{R}_{i,j} = \begin{cases} 1, & \text{for} \quad \vec{x}_i \approx \vec{x}_j \\ 0, & \text{for} \quad \vec{x}_i \not\approx \vec{x}_j \end{cases} \quad i, j = 1, \ldots, N \tag{1}$$

where N indicates the number of current states. $\vec{x}_i \approx \vec{x}_j$ means that they are approximately equal up to an error round defined as ϱ. In general, the recurrent matrix $\mathbf{R}$ compares the state and indicates the state similarity across the entire series [68]. The selection of similarity threshold ϱ is critical. Researchers have investigated the selection criterion in the literature based on the system states [69–71]. In Fig. 3, we choose $\varrho = 0.1$ for the visualization. Moreover, Fig. 3 suggest a typical vertical and horizontal lines pattern, which is suggested to be a laminar state or state idle case, while the sparse region indicates lower system complexity and vice versa [68].

Beyond the visual approach, recurrence quantification analysis (RQA) is also widely used to measure system-level complexity [70,72,

Table 2
Description of *selected* features recorded in the HITL simulations. Traffic density is directly obtained from the Metacraft. The latency variables are defined and collected following modified SWAT. Workload ratings are collected from the question probe. Latency measures and situational awareness questions are collected but not used in this work.

Feature names	Feature descriptions	Feature values
traffic_density	Total number of aircraft under ATC participant's control	Integer: 0–23
ready_latency	Time spent from screen appearing "Ready?" to participant pressing "Ready?" on the pop-up questionnaire.	Decimal: 0.01–60.00 s
query_latency	Time spent from pressing "Ready?" to selecting answers on the pop-up questionnaire	Decimal: 0.01–60.00 s
wl_rating	Workload rating select from the pop-up window	Integer: 1–7
sa_correct	Evaluation of selected responses on situation awareness questions from the pop-up window	0: no resp; 1: correct; 2: incorrect

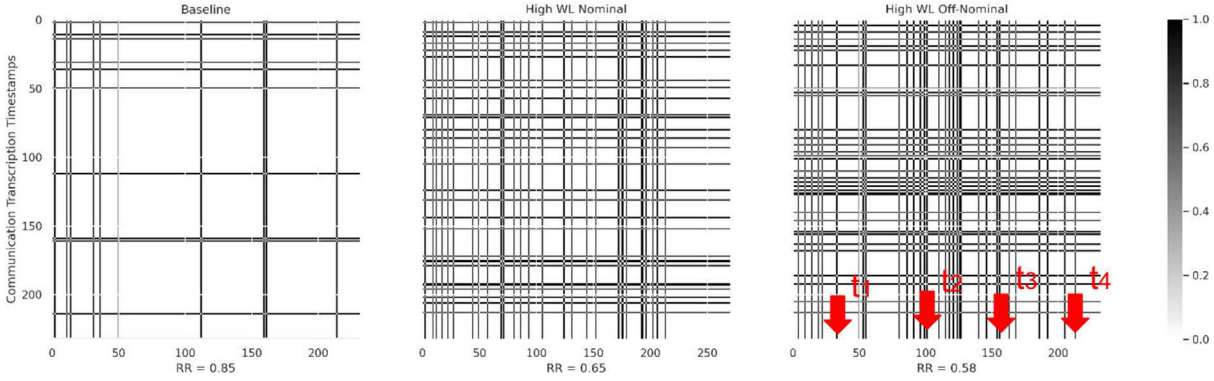
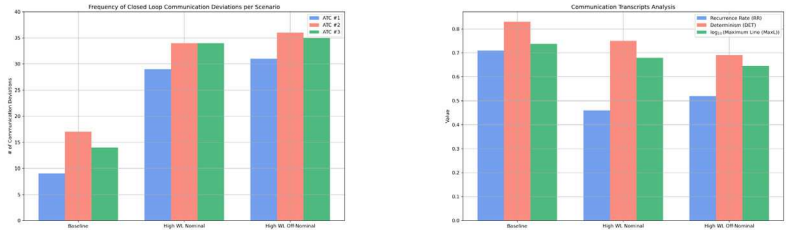


Fig. 3. Communication Transcription Visual Analysis: Recurrence Plot (RP). RP is used to quantify the overall tendency of recurrence in the system. Vertical/Horizontal lines indicate the laminar states do not change or change slowly over time [66].



(a) Number of Deviations that happened in each scenario. Showing first three ATCs.

(b) Recurrence Quantification Analysis (RQA).

Fig. 4. Communication data analysis on different scenarios. In (a), we show the histogram of the frequency of the communication deviations for each scenario (up to 3 ATCos). As discussed, the number of communication deviations indicates communication difficulties. In (b), we show the QRA under three different scenarios, showing the scenario complexities obtained from communications.

73]. Typically, RQA is based on the diagonal and vertical patterns of the RP. We use three types of complexity measures here.

Recurrence Rate (RR) is the simplest measure of RP, which is the averaged density of recurrence points in RP. RR represents the likelihood of a state returning to its ϱ-neighborhood in the phase space [74]. It is the measure of correlations between the ATCo and pilot communications.

Determinism (DET) refers to the degree of predictability or orderliness in a system's dynamics over time. A deterministic system is one in which future states can be precisely predicted from knowledge of the present state and the system's dynamics. In the context of recurrence plots, a high degree of DET is indicated by the presence of diagonal lines in the plot, which represent points in the system's trajectory that are close to each other in phase space and recur with a high degree of regularity. Conversely, a low degree of DET is indicated by a more random or chaotic pattern in the recurrence plot, with fewer or no diagonal lines. It indicates the predictability of ATCo and pilot interaction.

Maximum Line (MaxL) is a diagonal line that represents the longest connected sequence of recurrent points in the plot. It is the diagonal line that has the most points along it, and it indicates the most persistent pattern of recurrence in the system's dynamics. The length and frequency of maximum lines can provide insights into the regularity and predictability of the system's behavior over time. MaxL quantifies the stability of ATCo and pilot interaction.

Fig. 4 shows the histograms of communication deviations in (a), and the calculated measures of complexity in (b). As shown in Fig. 4(a), there is a notable rise in communication challenges from the baseline scenarios to the high workload scenarios. The prolonged occurrence of off-nominal events could contribute to a slightly heightened level of complexity. Additionally, the deviations in communication patterns can differ across Air Traffic Control Officers (ATCos), possibly due to variations in individual experience and seniority. Therefore, it can be concluded that both airspace density and off-nominal events contribute to an increase in communication complexity. As depicted in Fig. 4(b), it

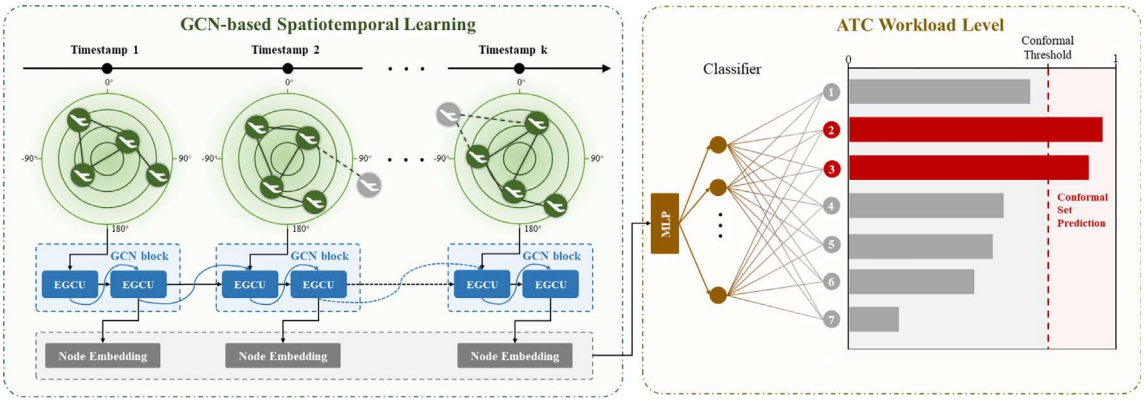


Fig. 5. Schematic illustration of conformalized EvolveGCN set prediction framework. We formulate the ATC workload level prediction as a time-series graph classification task, where each graph node represents *each aircraft under the ATC's control*. The number of nodes and weight (distance) of each edge can change across different timestamps. On the classifier side, we propose the conformal prediction set for improved ground truth coverage. Conformalization acts as a *post-hoc* procedure to post-process the prediction labels, where the softmax probability threshold is inferred on the calibration set.

is evident that the correlation, predictability, and stability of the system all exhibit a decrease from the baseline scenarios to the high workload scenarios. These findings provide valuable insights into understanding the behavior of different scenarios and guide our further studies on workload prediction.

4. Proposed ATC workload prediction framework

In this section, we describe the proposed workload machine learning prediction framework in Fig. 5. We first demonstrate the formulation and basic concepts of the graph learning problems in Section 4.1. The dynamical graph convolution learning algorithm and conformal prediction setup are explained in Section 4.2 and Section 4.3, respectively.

4.1. Problem formulation

As mentioned, the ATCo workload prediction task is viewed as a time-series dynamical graph classification task. In this paper, we use subscript $t \in \{1, \dots, T\}$ to demonstrate the timestamp and superscript $l \in \{1, \dots, L\}$ to denote the layer index. Graph neural networks (GNNs) are introduced to model the spatiotemporal layout of the airspace from the structured graph data with explicit message passing [75,76]. We denote a graph at timestamp t with vertices and edges, represented as $G_t = (V_t, E_t)$, where the number of nodes at timestamp t is $N_t^{nodes} = |V_t|$ and the number of edges at timestamp t is $N_t^{edges} = |E_t|$. The adjacency matrix $A_t \in \mathbb{R}^{N_t^{nodes} \times N_t^{nodes}}$. The constructed graph structure can be either directed or undirected depending on whether the edges are directed from one node to another. The dynamic graph is mentioned when the graph topology varies with time. Especially, the graphs in our work are undirected dynamical graphs. The constructed graph inputs are $A_t \in \mathbb{R}^{N_t^{nodes} \times N_t^{nodes}}$ and $X_t \in \mathbb{R}^{N_t^{nodes} \times N_t^{features}}$, where A_t is the adjacency matrix at each timestamp t and X_t is the node feature matrix.

Graph structure represents the spatial layout of the airspace. However, the dynamical graph constructed A_t is based on the geo-distance between two aircraft pairs, resulting in geospatial graphs on a two-dimensional space. This type of graph construction is widely adopted with acceptable complexity. Consequently, in this work, we adopt a scaling formula described in minimum-spanning-tree-based workload prediction task [26] to scale the horizontal and vertical distance between two aircraft pairs into one distance metric. Specifically, the graph is built by calculating the distance $\tilde{d_{ij}}$ between two aircraft pairs (i, j) at the same timestamp t. We use $d_{t,ij}$ to represent the horizontal distance and $h_{t,ij}$ to present the vertical separation distance between aircraft pairs (i, j). The scaling function is described in Eq. (2).

$$\tilde{d_{t,ij}} = \sqrt{d_{t,ij}^2 + s^2 h_{t,ij}^2} \tag{2}$$

where s is the spatial scaling factor which equalizes the separation on the horizontal and vertical dimension, as in Eq. (3).

$$s = \begin{cases} 0.005, & \text{for } \mathrm{alt}_i \le 29{,}000 \text{ and } \mathrm{alt}_j \le 29{,}000 \\ 0.0025, & \text{for } \mathrm{alt}_i > 29{,}000 \text{ or } \mathrm{alt}_j > 29{,}000 \end{cases} \quad i, j = 1, \dots, N \tag{3}$$

In this workload prediction task, we first obtain the constructed input X_t and A_t at each timestamp t. In such a way, we obtain the series of graphs $G_t, t \in \{1, \dots, T\}$. Then, we fill the workload ratings into another time series based on the self-evaluated workload rating during the HITL simulations, representing the prediction labels. Last, we use a moving window approach to build the correct input–output matching for supervised machine learning. The schematic illustration of the moving window process is shown in Fig. 6. We define a window of size κ and move the window along the time axis of the series of inputs, with a stride of 1. The time-series graph of size κ is denoted as $\{G_t\}_\kappa$. Similarly, we move the window function along the prediction labels and obtain the workload ratings by claiming the last reported workload value within the current window. Then, if Y_t denotes the ground truth workload label at timestamp t, we mathematically formulate the problem into,

$$Y_t = \mathrm{EvolveGCN}(\{G_t\}_\kappa) \tag{4}$$

4.2. Evolving graph convolution network

Spatial graph convolutional networks (GCN) [77] convolve the input A_t and X_t using the derived compact form,

$$H_t^{l+1} = \sigma(\hat{A}_t H_t^l W_t^l), \quad with \quad H_t^0 = X_t \tag{5}$$

where σ is the activation function (i.e., ReLU). $\hat{A}_t$ is a normalized version of A_t, to account of numerical instability. Specially, $\hat{A}_t$ is defined as, $\hat{A}_t = \tilde{D}_t^{-\frac{1}{2}} \tilde{A}_t \tilde{D}_t^{-\frac{1}{2}}$, $\tilde{A}_t = A_t + I_t$, $\tilde{D}_{t,ii} = \sum_j \tilde{A}_{t,ij}$. It is clear that H_t has the same dimension as X_t as $H_t \in \mathbb{R}^{N_t^{nodes} \times N_t^{features}}$. $W_t \in \mathbb{R}^{N_{features} \times N_{featrues}}$ is the kernel parameters. For multiple graph convolutional layer setup, H_t^{l+1} stands for the updated graph embedding of convolutional layer $l+1$ at timestamp t. Specifically, in classification problems, the activation function σ at the output layer L is the softmax function.

Evolving Graph Convolution Network (EvolveGCN) improves GCN by introducing recurrence layers to capture the dynamism underlying a time-series graph. Two types of EvolveGCN are presented [78], depending on the recurrent updating architecture.

The first variant treats the GCN kernel parameter W_t^l as the hidden state of recurrent learning function and updates W_t^l with a gated

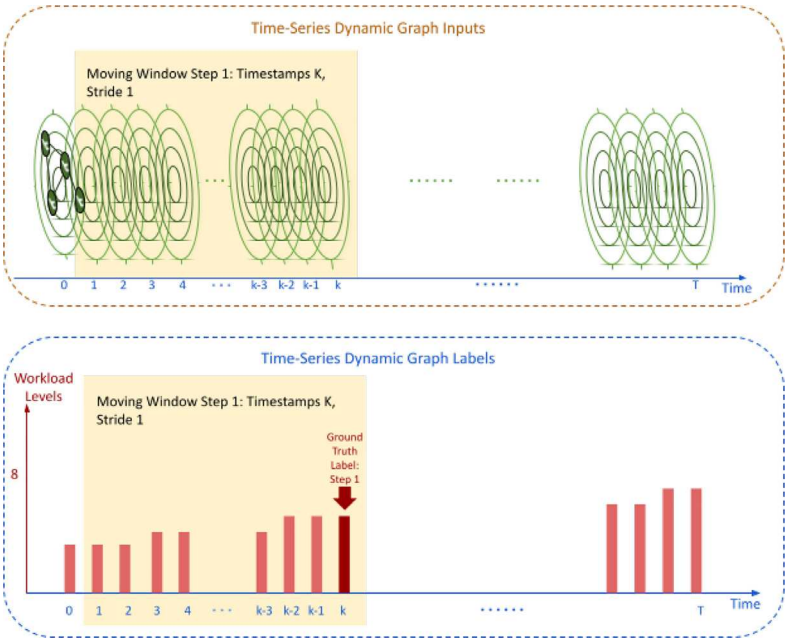


Fig. 6. Schematic illustration of the moving window approach. At each step, the moving window moves 1 timestamp (5s) along the temporal dimension (stride 1). Each series of graphs contains a graph of κ timestamps. The workload ground truth label for the graph series input is the workload level (1–7) reported at the last timestamp, collected from the human-in-the-loop experiment. This setup allows the model to capture long-term spatial relationships and result in a prediction at every timestamp since κ. For abbreviation, the graph input is represented by a radar plot.

recurrent unit (GRU), while the node embeddings of node features are still contained within the GCN hidden state tensor H_t^l. EvolveGCN-H is used to denote this variant Eq. (6). This requires a special design of GRU computation flows as described in [78].

$$W_t^l = GRU(H_t^l, W_{t-1}^l) \tag{6}$$

$$H_t^{l+1} = \sigma(\hat{A}_t H_t^l W_t^l) \tag{7}$$

Another variant of EvolveGCN is the -O version Eq. (8), where the kernel parameter W_t^l is treated as input of recurrent learning without considering the temporal correlations between node embeddings. The implementation of EvolveGCN-O is straightforward by extending dimensions.

$$W_t^l = LSTM(W_{t-1}^l) \tag{8}$$

$$H_t^{l+1} = \sigma(\hat{A}_t H_t^l W_t^l) \tag{9}$$

Either an EvolveGCN-H or EvolveGCN-O is denoted as an Evolving Graph Convolution Unit (EGCU), as shown in Fig. 5. In both ways, the EGCU first updates the GCN weights and then propagates the hidden states through the layers. Several layers of EGCU form a GCN block in Fig. 5. For a graph learning problem with large feature space, EvolveGCN-H is more effective since the feature embedding recurrence is also considered. Otherwise, EvolveGCN-O is more focused on learning the graph topology structure changes.

4.3. Conformal prediction

In this section, we provide a brief overview of conformal prediction (CP). For the classification task mentioned above, we have EvolveGCN acted as the classifier C, which outputs an estimated probability for each class, i.e., $p \in [0,1)^{\varkappa}$ for $\varkappa$ classes. We reserve a small amount of data called *calibration set* to calculate the probability score threshold $\hat{q}$ such that the following condition holds on the test set,

$$1-\alpha \le \mathbb{P}(Y_{test} \in C(X_{test})) \le 1-\alpha+\frac{1}{n+1} \tag{10}$$

where the test dataset is the unused test set to evaluate model performance. $\alpha \in (0,1]$ is the pre-defined tolerated error rate. This is to guarantee that the model is $1-\alpha$ confident that the model prediction set contains the correct ground truth label. This equation is also known as the marginal conformal coverage guarantee, which has been proved in the literature [79,80]. Notably, the *calibration* is the key step to find $\hat{q}$. Suppose we define the concept of the conformal score by one minus the softmax probability of the true class, $\hat{q}$ is defined to be the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the conformal scores. $\lceil \rceil$ is the ceiling function to correct the quantile. Then, the prediction of a new test sample will be all classes with a softmax score higher than $\hat{q}$. The prediction set will be larger if the model is uncertain about the prediction labels or if the input is out-of-distribution. Intrinsically, the size of the prediction set is the indicator of model uncertainty.

Conformal prediction (CP) has been studied from various angles by researchers. However, the classical CP method is susceptible to coverage issues due to its tendency to produce the smallest average size of prediction sets [81]. Specifically, CP tends to overcover hard data samples while undercover simple ones. To address this issue, researchers have proposed an approach called adaptive conformal prediction [82,83]. The underlying principle of this method is to compute the conformal threshold $\hat{q}$ based on the cumulative softmax score across $\varkappa$ classes.

It is also noteworthy to discuss conformal evaluation methods, which are adopted in evaluating our model. To determine the model's performance, a straightforward method is to examine the histogram of prediction set sizes visually. Essentially, a larger size of the prediction set implies that the model is facing certain data quality problems, while the variation in the set size can provide insights into the model's ability to differentiate between easy and difficult input samples.

$$\mathbb{P}[Y_{test} \in C(X_{test}) | X_{test}] \ge 1-\alpha \tag{11}$$

Conditional coverage is a feasible approach to evaluate the adaptivity of conformal prediction. For instance, in a classification setting, we seek to find the prediction sets with exactly $1-\alpha$ coverage for any input data sample, as in Eq. (11). The conditional coverage concept is a stronger metric than the marginal coverage mentioned above. Some literature mentioned that conditional coverage is impossible to achieve in most general cases [84]. **Size-stratified coverage (SSC)**

Table 3
List of fine-tuned model parameters used in EvolveGCN training under three different simulation scenarios.

	Baseline	High Workload Nominal	High Workload Off-Nominal
Number of EGCU Layers	2	2	4
EGCU Layer Dimensions	64	128	64
Dropout Ratio	0.25	0.5	0.25
Learning Rate	0.001	0.0015	0.0005

Table 4
Workload Level Prediction: Comparison between different workload prediction methods.

ATC workload level prediction	Baseline		High workload nominal		High workload off-nominal	
	MicroF1	MacroF1	MicroF1	MacroF1	MicroF1	MacroF1
Simple LRw/ Density	0.306	0.218	0.307	0.198	0.323	0.195
Simple LRw/ Graph Feature	0.331	0.236	0.383	0.263	0.350	0.255
2-layer MLP	0.364	**0.283**	0.455	0.386	0.459	0.367
GCN	0.404	0.218	0.580	0.401	0.526	0.352
EvolveGCN-O	**0.545**	0.277	**0.740**	**0.632**	**0.695**	**0.472**
EvolveGCN-H	0.413	0.221	0.593	0.474	0.581	0.414

metric is a general metric to evaluate how close the model is able to achieve Eq. (11). SSC metric is a way to evaluate the performance of conformal prediction models. It is based on the idea that prediction sets of different sizes may have different properties and should be evaluated separately. The key is to group test samples into different size strata based on the size of their prediction sets and compute the averaged empirical coverage on each size strata. SSC metric can be useful to diagnose specific issues, such as overcoverage or undercoverage, that may be related to the size of the prediction sets. In addition, another conformal prediction evaluation method has been proposed recently [85], where a calibration plot of the prediction error versus the specified significance level (α) is used. Remarkably, In Section 5, we evaluate our model with these metrics.

CP provides a rigorous way to measure the uncertainty associated with the predictions made by a machine learning model and to express this uncertainty in the form of prediction intervals or regions that can be used to guide decision-making. This can be particularly useful in critical engineering applications where accurate prediction intervals or regions are essential [86]. CP has been widely adopted in drug discovery [87], medical diagnosis [88], and robotics [89]. CP is a unified *post-hoc* softmax score calibration process to generate prediction sets for any classification model [90–92]. In this work, we propose to use CP for aviation decision support. Specifically, the prediction set comes from CP gives uncertain prediction label suggestions, i.e., workload rating of 3, 5, 7. While the isolated classification label in workload prediction is not reasonable, we propose to fill the intermediate workload ratings based on the minimum predicted rating and the maximum predicted ratings.

5. Experiments

In previous sections, we have introduced the HITL data collection process, the problem definition, the machine learning model system design, and conformal prediction for better ground truth label coverage Section 4. In this section, we present a comprehensive of experiments to test and evaluate the proposed model. We first discuss several evaluation metrics used for this classification task. Then we report the classification accuracy from the machine learning model with a few implementation details. Lastly, conformal prediction set results are reported with several conformal coverage evaluation methods mentioned in Section 4.3. The validation set is used to tune hyper-parameters, and the testing dataset results are evaluated based on the best validation epoch.

5.1. Evaluation metrics

The F-score, or F-measure, is a binary classification metric used in statistical analysis to assess the accuracy of test samples. Specifically, the F1-score is defined as the symmetrical harmonic mean of precision and recall [93]. F1-score can also be used for multi-class classification by taking either the micro-averaging (MicroF1) or the macro-averaging (MacroF1).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$\text{MicroF1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{14}$$

Eq. (12) gives the mathematical formulation of MicroF1, where each sample is considered independently without considering which class this sample belongs to. MicroF1 treats each data input equally but is biased on class frequency. MicroF1 is useful when the classification task is unbalanced, meaning that some classes have many more instances than others. In this case, the MicroF1 gives equal weight to each instance, regardless of its class.

$$\text{MacroF1} = \frac{\text{F1}_1 + \text{F1}_2 \cdots, +\text{F1}_n}{\text{n}} \tag{15}$$

On the contrary, MacroF1 average the F1 score across all classes as in , where n is the number of classes, and $\text{F1}_1 + \text{F1}_2 \cdots, +\text{F1}_n$ are the F1 scores for each class. MacroF1 treats each class equally regardless of the size of samples within each class thus, it is biased on the number of samples. MacroF1 is useful when the classification task is balanced, meaning that each class has approximately the same number of instances. In this case, MacroF1 gives equal weight to each class, regardless of its frequency.

In this work, we are encountering a highly-imbalanced classification problem Fig. 2. Thus, MicroF1 is a better indicator than MacroF1. It is noteworthy to mention that n in should be adjusted to exclude the class labels that are not presented in either ground truth or predictions. We report both metrics in our studies.

5.2. Implementation details

Although the commonly adopted node-level or link-level classification objective is prevalent, the proposed workload prediction framework is instead a graph-level classification task [78]. Thus, on the output layer, we take the aggregated class probability score across each node to get a unified score of the entire graph. We adopted the grid-search strategy to search for key parameters and fine-tuned the

neural network model with the data collected from three different scenarios. The key parameters used are the number of EvolveGCN layers (EGCU layers), and the dimensions of these layers. Dropout is used in the classifier to address overfitting. Table 3 lists the fine-tuned key model parameters under three simulation scenarios. Other parameters, such as the dimension of classifiers, are kept the same as the original implementation [78]. Parameter tuning on the classifiers might be useful, but it is beyond the scope of this study.

Moreover, as discussed in Section 3, the first reported workload rating starts at 3 min of the 25 min duration. This corresponds to the $36th$ timestamp with $5s$ interval in the collected flight traffic data. Consequently, the moving window size κ in our experiment is 36, with a stride of 1. We separate the data into train, validation, and test sets with a ratio of $[0.4, 0.3, 0.3]$. The validation set is used for deep learning model hyper-parameter tuning. Moreover, the validation set is also used as the calibration set to find the CP threshold $\hat{q}$.

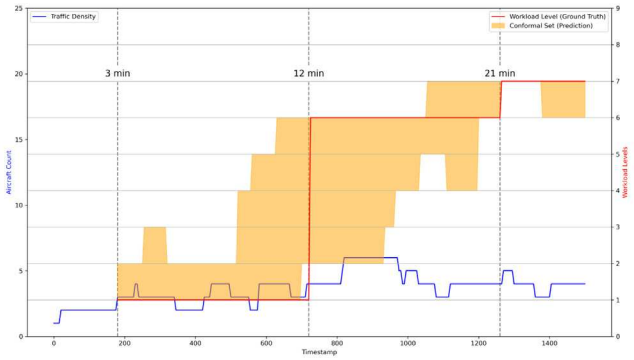
5.3. Experiment results

In this work, we compare our model with both classical methods (i.e., linear regressions (LRs)) and simple data-driven learning methods (i.e., fully-connected neural networks/multilayer perceptrons (MLPs)). In [23], the authors also conducted high-fidelity human-in-the-loop simulations to study the impact of traffic density features on controller workload. They found that the workload rating of the enroute center controller is proportional to the number of aircraft with a slope of 0.306 and bias of −3.373. They also identified the primary sources of workload for controllers, including airspace and traffic management, communication, and coordination tasks with workload management suggestions. In [26,28], the authors create graph-structured airspace data structure – minimum-spanning trees but propose several handcraft features based on the histogram of node features. Then a two-layer fully-connected neural network is used for prediction based on handcrafted features and shows remarkable performance. However, the workload ratings are directly generated from the traffic density, where thresholds of 7 and 17 separate workload ratings into low, medium, and high scenarios. Likewise, inspired by this work, our proposed method adopts a graph structure to represent the spatiotemporal layout. We utilize the recent advancement in graph learning and learn from the graph structure without handcrafted features.

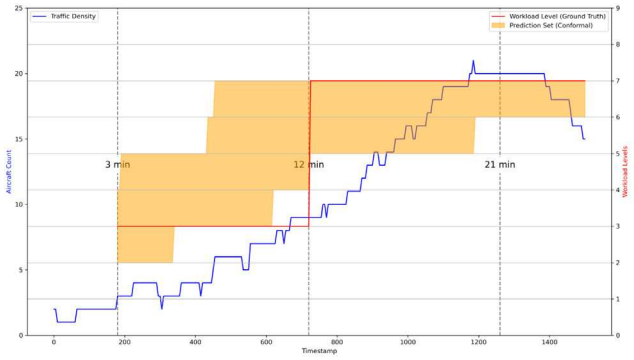
In Table 4, comparing the first two rows, we first show that including additional graph node features can achieve higher prediction accuracy, even for simple LRs. Despite the traffic density features, additional graph node features are traffic conflict features (i.e., horizontal/vertical minimum separation to nearby aircraft). For the MLP with handcraft features, we generate second-order statistics of the sum and difference histograms introduced in [26]. As a reference to EvolveGCN, we also conduct an experiment on vanilla GCN. This can be easily achieved by removing the LSTM layer in Eq. (8). We show that EvolveGCN can achieve significantly higher MicroF1 and MacroF1 than LP, MLP, and GCN. Moreover, the -O variant EvolveGCN outperforms the -H variant. One of the major reasons is the selection of top K indices reduces the hidden dimension of EGCU, due to the low node feature dimensions and a small number of nodes in graphs (i.e., only one aircraft showing up at the first timestamp) at certain timestamps.

5.4. Conformal prediction results

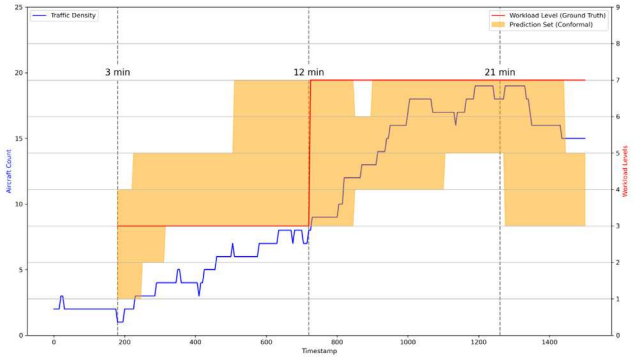
As mentioned, we use conformal prediction to improve prediction accuracy further. In Fig. 7, the prediction on one test participant is shown. The conformal prediction set coverage is the shaded region. The conformal prediction is generated with a tolerated error rate of $\alpha = 5\%$. The blue solid lines show the real-time aircraft density (left axis) in the simulation, and the red lines are the interpolated workload ratings. At 3 min, 12 min, and 21 min, the participants are required to submit their workload rating to the computer. The ground truth workload ratings are colored in red (right axis). The conformal prediction set covers most of the ground truth but is undercover at several spots. As discussed in Section 4.3, we further evaluate the conformal prediction coverage.



(a) Baseline Condition



(b) High Workload Nominal Condition



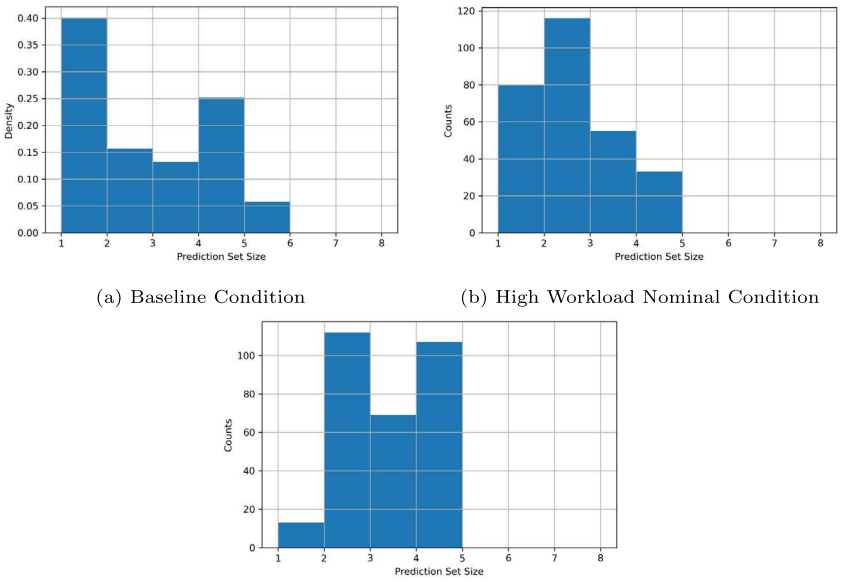
(c) High Workload Off-Nominal Condition

Fig. 7. Visualization of conformal predictions on the test sample. For the workload level prediction task, we set our prediction as the range between the lowest predicted workload level and the highest predicted workload level.

5.5. Conformal coverage evaluation

We adopt different conformal coverage evaluation metrics to examine the performance of our conformal set. Firstly, we plot the histogram of set sizes. A high average set size suggests that the conformal prediction procedure is imprecise, which could indicate issues with the score or underlying model. Secondly, the range of set sizes indicates whether the prediction sets adapt properly to the complexity of examples. A wider range is typically preferred because it implies that the procedure accurately differentiates between simple and challenging inputs. We show the histograms in Fig. 8. The size of conformal prediction is typically around 5. The spread for baseline and high workload nominal conditions looks reasonable. The model is able to distinguish hard and easy samples. However, the spread for high workload off-nominal conditions indicates potential scoring issues or simply difficult data [80].

Following the discussion in Section 4.3, we investigate the size-stratified coverage (SSC) to evaluate the condition coverage and plot



(a) Baseline Condition (b) High Workload Nominal Condition

(c) High Workload Off-Nominal Condition

Fig. 8. Histogram of set sizes on test set predictions. The spread of the histogram shows the difficulty of making a correct prediction.

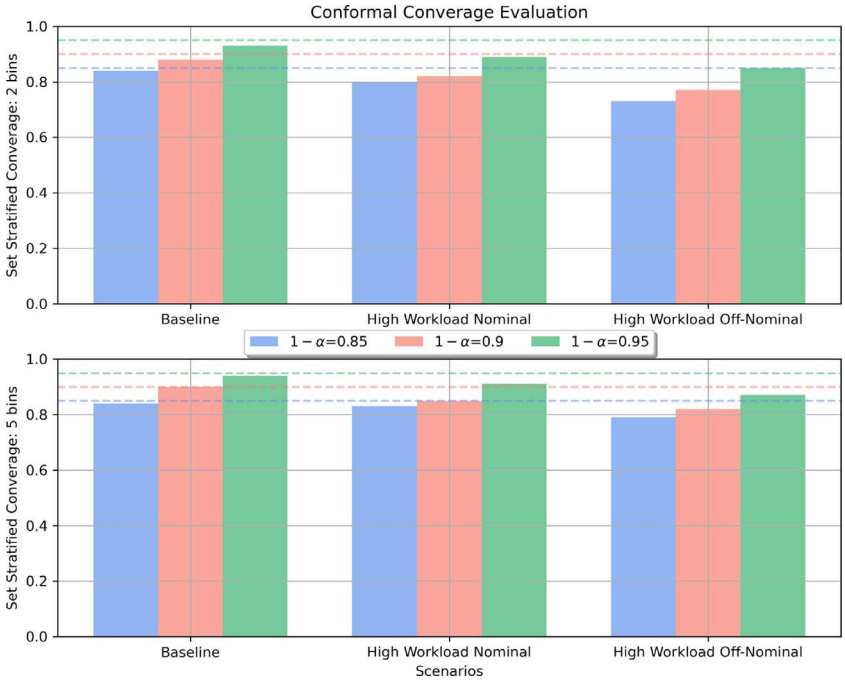
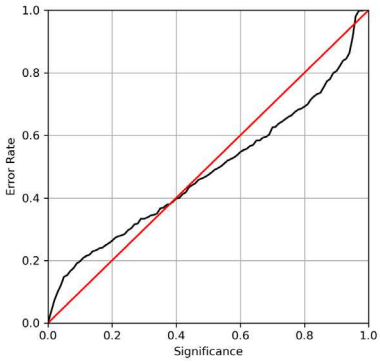


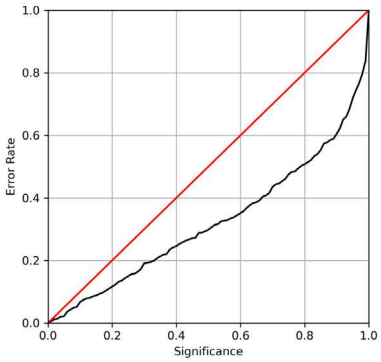
Fig. 9. Conformal coverage evaluation with various desired α values under three workload simulation conditions. We use the Size-Stratified coverage (SSC) metric better to represent the adaptive coverage of the conformal set coverage.

the figures in Fig. 9. The dash lines are desired coverage values. In this figure, we consider three desired error rates, $\alpha = 0.15, 0.1, 0.05$, for three scenarios. We use two possible $C(x)$ cardinalities of two bins and five bins. In other words, we divide the predicted sets into different size categories (e.g., sets of size 2, sets of size 5.) and calculate the percentage of times that the true value falls within each category. We discover that the prediction coverage of baseline conditions shows a good sign, but the two high workload scenarios tend to under-coverage. Again, the reasons still come from the unsatisfactory of the collected data, which leads to lower reported F1 values.

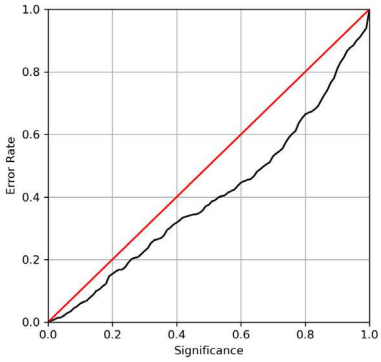
To further look at the coverages, we adopt another recently proposed figure to better show the prediction coverage violations [85]. In Fig. 10, we show three figures for three simulation scenarios. The x-axis shows the tolerated error rate (the specified significant level), and the y-axis shows the fractions of failed prediction sets (the number of prediction samples where the ground truth label is not in the conformal



(a) Baseline Condition



(b) High Workload Nominal Condition



(c) High Workload Off-Nominal Condition

Fig. 10. The calibration plot illustrates the observed prediction error, which is the proportion of true labels that are not included in the prediction set, plotted against the pre-specified significance level ϵ, or the tolerated error rate. The conformal predictor is deemed valid only when the observed error rate is within the limit of ϵ, i.e., the observed error rate should align closely with the diagonal line representing the tolerated error rate for all significance levels. One of the key advantages of conformal predictors is their ability to offer valid predictions even when new examples are independently and identically distributed with the training examples. Additionally, we use Kolmogorov–Smirnov (K-S) test to test the distributions of predictions in the calibration set and test data under three conditions. In the K-S test, the null hypothesis is that the calibration and test samples are drawn from the same data distribution. The corresponding p-values obtained for three conditions are, (a) $1.98e-12$; (b) $3.68e-48$; (c) $2.18e-55$. All three values below 5%, are considered two-sided statistically significant. Thus, the hypothesis holds true.

prediction set). This property holds true in high workload scenarios but not in the baseline scenarios when the significance level is lower than 0.4 or higher than 0.95.

6. Conclusions

In this paper, we investigate the workload prediction problem. We formulate the problem into a time-series dynamic graph classification task with changing graph topologies. We demonstrate the effectiveness of this proposed method from real-world human-in-the-loop air traffic control simulations, in which participants are retired air traffic controllers. We show that traffic density features and traffic conflict features have a positive influence on workload predictions. Algorithm-wise, the graph-structured data-driven learning model outperforms the existing practices in workload prediction research literature (i.e., simple regressions, simple neural networks with handcrafted features).

6.1. Limitations

There are several limitations. Firstly, we only have limited resources to conduct the HITL experiments in a simulation environment. Real-world scenarios can be immensely different from simulation scenarios, with either fewer or more deviations. Secondly, data quality is critical for developing a successful machine-learning algorithm. In this work, we have to use the corrected workload rating data due to the poor quality of the originally collected workload ratings, with only six retired ATCo participants [62]. Another critical part is modeling the different ATC strategies adopted by different controllers, which can result in a multi-modal machine learning setup. The benefit of including ATCo strategies has also been discussed in the literature [4]. Lastly, better algorithm development can help with improved workload prediction performance. The single prediction label made by either EGCU-O or EGCU-H can be improved, despite the data quality issue.

6.2. Insights

This work is beneficial for the non-intrusive, uninterrupted executive controller workload prediction, and it is purely based on the flight traffic data. First, we show that both traffic density and traffic conflict features contribute to higher prediction accuracy. Then, we show that model of the spatiotemporal airspace layout as a dynamic time-series graph learning problem has great potential for ATC workload level predictions. Additionally, we explore the possibility of further accuracy improvement by introducing a *post-hoc* classification score processing process, namely conformal prediction, which can be used to generate multiple classification labels adaptively.

Based on these insights, we propose several research directions that might be interesting to researchers,

- We are expecting a significant performance improvement by conducting more HITL simulations or real-world ATC experiments, collecting additional high-quality data, and data-driven model refinement. Spatiotemporal graph learning is a popular theoretical

research direction, and better graph learning model architecture is expected, which results in better workload prediction performances.

- The window function setup for input–output data matching is flexible for any practical requirement in the real world. The length of the window determines the history length to be considered, while the stride size defines the prediction horizon.
- The workload prediction problem formulation can be alternated from workload rating classification to workload rating regression task. This setup is algorithm-wise more reasonable for uncertainty calibration with conformal prediction but requires significant experiment setup change. For instance, the current rating-based question prob in modified SWAT and NASA TLX will be modified to continuous variables. A considerable modification of the HITL simulations is desired.
- There are still several important questions that remain unanswered, such as how to incorporate real-time data and feedback into the prediction model and how to adapt the model to different types of air traffic control systems. Future research in this area could also explore the impact of other factors, such as weather conditions and aircraft type, on controller workload and safety.
- Several works of literature quires the validity of only using traffic-related factors to predict mental workload, where a significant part of pilot-controller interactions and feedback are missing [4]. In further studies, we propose to build a predictive model that can consider reciprocal feedback interactions (i.e., the communication deviations [7] in Section 3) from a learning perspective.
- Further studies can also combine trajectory prediction models such that the task demands can be predicted first and then perform workload forecast in real-time. In such a way, either deterministic or probabilistic trajectory prediction models can act as moderators of workload models [31,94–96].

We believe the above discussions have practical implications for aviation authorities, airlines, and air traffic management providers. Specifically, our workload prediction model could be used to inform scheduling and staffing decisions, optimize resource allocation, and support proactive safety management. However, it is important to note that successfully implementing such interventions will require collaboration and communication across stakeholders.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Yongming Liu reports financial support was provided by NASA.

Data availability

Data will be made available on request.

Acknowledgments

The research reported in this paper was supported by funds from NASA University Leadership Initiative program, USA (Contract No. NNX17AJ86 A, PI: Yongming Liu, Technical Officer: Anupa Bajwa). The support is gratefully acknowledged.

References

[1] P.A. Hancock, N. Meshkati, Human Mental Workload, North-Holland Amsterdam, 1988.
[2] T.B. Sheridan, T.B. Sheridan, K. Maschinenbauingenieur, T.B. Sheridan, T.B. Sheridan, Humans and Automation: System Design and Research Issues, Vol. 280, Human Factors and Ergonomics Society Santa Monica, CA, 2002.
[3] F. Nachreiner, P. Nickel, I. Meyer, Human factors in process control systems: The design of human–machine interfaces, Saf. Sci. 44 (1) (2006) 5–26.
[4] 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 (3) (2007) 376–399.
[5] D. Gopher, E. Donchin, Workload: An Examination of the Concept, John Wiley & Sons, 1986.
[6] D. Gianazza, Forecasting workload and airspace configuration with neural networks and tree search methods, Artif. Intell. 174 (7–8) (2010) 530–549.
[7] J. Djokic, B. Lorenz, H. Fricke, Air traffic control complexity as workload driver, Transp. Res. C 18 (6) (2010) 930–936.
[8] G. Tobaruela, W. Schuster, A. Majumdar, W.Y. Ochieng, L. Martinez, P. Hendrickx, A method to estimate air traffic controller mental workload based on traffic clearances, J. Air Transp. Manag. 39 (2014) 59–71.
[9] H. Wang, D. Gong, R. Wen, Air traffic controllers workload forecasting method based on neural network, in: The 27th Chinese Control and Decision Conference, 2015 CCDC, IEEE, 2015, pp. 2460–2463.
[10] FAA, Air traffic by the numbers, 2020, https://www.faa.gov/air_traffic/by_the_numbers/media/Air_Traffic_by_the_Numbers_2020.pdf.
[11] S.V. Ligda, M.L. Seeds, M.J. Harris, C.S. Lieber, M. Demir, N. Cooke, Monitoring human performance in real-time for NAS safety prognostics, in: AIAA Aviation 2019 Forum, 2019, p. 3411.
[12] I. Dhief, Z. Wang, M. Liang, S. Alam, M. Schultz, D. Delahaye, Predicting aircraft landing time in extended-TMA using machine learning methods, in: ICRAT 2020, 9th International Conference for Research in Air Transportation, 2020.
[13] C.S. Lieber, M. Demir, N. 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, 2021, p. 2320.
[14] B. Hilburn, Cognitive complexity in air traffic control: A literature review, EEC Note 4 (04) (2004) 1–80.
[15] Federal Aviation Regulations, Title 14-Aeronautics and Space, US Government Printing Office, Washington, USA, 2017.
[16] National Transportation Safety Board, Annual Review of Aircraft Accident Data, US Air Carrier Operations: Calendar Year 2001, Rep. No. ARC-06–01, Author Washington, DC, 2001.
[17] D.-T. Pham, S. Alam, V. Duong, An air traffic controller action extraction-prediction model using machine learning approach, Complexity 2020 (2020) 1–19.
[18] Y. Heng, M. Wu, X. Wen, et al., Identifying key risk factors in air traffic controller workload by SEIR model, Math. Probl. Eng. 2022 (2022).
[19] R. Xiong, Y. Wang, P. Tang, N.J. Cooke, S.V. Ligda, C.S. Lieber, Y. Liu, Predicting separation errors of air traffic controllers through integrated sequence analysis of multimodal behaviour indicators, Adv. Eng. Inform. 55 (2023) 101894.
[20] EASA, EASA concept paper: Artificial intelligence roadmap: A human-centric approach to AI in aviation, 2021, https://www.easa.europa.eu/en/downloads/109668/en. (Accessed 20 March 2023).
[21] EASA, EASA concept paper: First usable guidance for level 1&2 machine learning applications: A deliverable of the EASA AI roadmap, 2023, https://www.easa.europa.eu/en/downloads/137631/en. (Accessed 20 March 2023).
[22] C.A. Manning, S.H. Mills, C.M. Fox, E.M. Pfleiderer, H.J. Mogilka, Using Air Traffic Control Taskload Measures and Communication Events to Predict Subjective Workload, Tech. Rep., FEDERAL AVIATION ADMINISTRATION OKLAHOMA CITY OK CIVIL AEROMEDICAL INST, 2002.
[23] S. Hah, B. Willems, R. Phillips, The effect of air traffic increase on controller workload, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 50, No. 1, SAGE Publications Sage CA, Los Angeles, CA, 2006, pp. 50–54.
[24] J. Crutchfield, C. Rosenberg, Predicting Subjective Workload Ratings: A Comparison and Synthesis of Operational and Theoretical Models, Tech. Rep., FEDERAL AVIATION ADMINISTRATION OKLAHOMA CITY OK CIVIL AEROMEDICAL INST, 2007.
[25] T. Edwards, S. Sharples, J.R. Wilson, B. Kirwan, Factor interaction influences on human performance in air traffic control: The need for a multifactorial model, Work 41 (Supplement 1) (2012) 159–166.
[26] B. Sridhar, K.S. Sheth, S. Grabbe, Airspace complexity and its application in air traffic management, in: 2nd USA/Europe Air Traffic Management R&D Seminar, Federal Aviation Administration Washington, DC, 1998, pp. 1–6.
[27] G.B. Chatterji, B. Sridhar, Neural network based air traffic controller workload prediction, in: Proceedings of the 1999 American Control Conference (Cat. No. 99CH36251), Vol. 4, IEEE, 1999, pp. 2620–2624.
[28] G. Chatterji, B. Sridhar, Measures for air traffic controller workload prediction, in: 1st AIAA, Aircraft, Technology Integration, and Operations Forum, 2001, p. 5242.
[29] A. Majumdar, W.Y. Ochieng, Factors affecting air traffic controller workload: Multivariate analysis based on simulation modeling of controller workload, Transp. Res. Rec. 1788 (1) (2002) 58–69.
[30] T. Edwards, L. Martin, N. Bienert, J. Mercer, The relationship between workload and performance in air traffic control: exploring the influence of levels of automation and variation in task demand, in: Human Mental Workload: Models and Applications: First International Symposium, H-WORKLOAD 2017, Dublin, Ireland, June 28-30, 2017, Revised Selected Papers 1, Springer, 2017, pp. 120–139.

[31] S.C. Corver, D. Unger, G. Grote, Predicting air traffic controller workload: trajectory uncertainty as the moderator of the indirect effect of traffic density on controller workload through traffic conflict, Hum. Factors 58 (4) (2016) 560–573.

[32] K. Sharma, H. Iyer, R. Pant, Cognitive ability criterion for expertise in air traffic control task, in: AIAA SCITECH 2022 Forum, 2022, p. 2449.

[33] L.L. Di Stasi, M. Marchitto, A. Antolí, T. Baccino, J.J. Cañas, Approximation of on-line mental workload index in ATC simulated multitasks, J. Air Transp. Manag. 16 (6) (2010) 330–333.

[34] H.A. Abbass, J. Tang, R. Amin, M. Ellejmi, S. Kirby, Augmented cognition using real-time EEG-based adaptive strategies for air traffic control, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 58, No. 1, SAGE Publications Sage CA, Los Angeles, CA, 2014, pp. 230–234.

[35] P. Aricò, G. Borghini, G. Di Flumeri, A. Colosimo, S. Bonelli, A. Golfetti, S. Pozzi, J.-P. Imbert, G. Granger, R. Benhacene, et al., Adaptive automation triggered by EEG-based mental workload index: a passive brain-computer interface application in realistic air traffic control environment, Front. Hum. Neurosci. 10 (2016) 539.

[36] B. Hilburn, G. Flynn, Toward a non-linear approach to modeling air traffic complexity, in: 2nd Human Performance Situation Awareness and Automation Conference, 2004.

[37] F.T. Durso, A.L. Alexander, Managing workload, performance, and situation awareness in aviation systems, in: Human Factors in Aviation, Elsevier, 2010, pp. 217–247.

[38] R.M. Rose, C.D. Jenkins, M.W. Hurst, et al., Air Traffic Controller Health Change Study: A Prospective Investigation of Physical, Psychological and Work-Related Changes, Tech. Rep., Civil Aerospace Medical Institute, 1978.

[39] P. Kopardekar, S. Magyarits, Measurement and prediction of dynamic density, in: Proceedings of the 5th Usa/Europe Air Traffic Management R & D Seminar, Vol. 139, 2003.

[40] D. Delahaye, S. Puechmorel, Air traffic complexity: Towards an intrinsic metric, in: Proceeding of the 3rd USA/Europe Air Traffic Management R and D Seminar, 2000.

[41] D. Gianazza, K. Guittet, Selection and evaluation of air traffic complexity metrics, in: 2006 Ieee/Aiaa 25TH Digital Avionics Systems Conference, IEEE, 2006, pp. 1–12.

[42] R.H. Mogford, J. Guttman, S. Morrow, P. Kopardekar, The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature, CTA INC MCKEE CITY NJ, 1995.

[43] K. Kallus, D. Van Damme, A. Dittman, Integrated Job and Task Analysis of Air Traffic Controllers: Phase 2, Task Analysis of En-Route Controllers (European Air Traffic Management Programme Rep. No. HUM. ET1. ST01. 1000-REP-04), EUROCONTROL, Brussels, Belgium, 1999.

[44] B. Kirwan, R. Scaife, R. Kennedy, Investigating complexity factors in UK air traffic management, Hum. Factors Aerosp. Saf. 1 (2) (2001).

[45] J.M. Histon, R.J. Hansman, G. Aigoin, D. Delahaye, S. Puechmorel, Introducing structural considerations into complexity metrics, Air Traffic Control Q. 10 (2) (2002) 115–130.

[46] 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 (5) (2009) 638–651.

[47] 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 (1) (2011) 127–140.

[48] J.H. Crump, Review of stress in air traffic control: Its measurement and effects, Aviat. Space Environ. Med. (1979).

[49] 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 (4) (2006) 297–314.

[50] F. Trapsilawati, M.K. Herliansyah, A.S.A.N.S. Nugraheni, M.P. 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 (3) (2020) 678–696.

[51] 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. Inform. 42 (2019) 100940.

[52] 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. Inform. 41 (2019) 100919.

[53] Y. Liang, S. Samtani, B. Guo, Z. Yu, Behavioral biometrics for continuous authentication in the internet-of-things era: An artificial intelligence perspective, IEEE Internet Things J. 7 (9) (2020) 9128–9143.

[54] C. Berghoff, M. Neu, A. von Twickel, The interplay of AI and biometrics: Challenges and opportunities, Computer 54 (9) (2021) 80–85.

[55] A.J. Masalonis, M.B. Callaham, C.R. Wanke, Dynamic density and complexity metrics for realtime traffic flow management, in: Proceedings of the 5th USA/Europe Air Traffic Management R & D Seminar, Budapest, Hungary, 2003, p. 139.

[56] D. Knorr, L. Walter, Trajectory uncertainty and the impact on sector complexity and workload, SESAR Innov. Days 29 (2011).

[57] D. Gianazza, Learning air traffic controller workload from past sector operations, in: ATM Seminar, 12th USA/Europe Air Traffic Management R&D Seminar, 2017.

[58] Y. Liu, K. Goebel, Information fusion for national airspace system prognostics: A NASA ULI project, in: Proceedings of the 10th Annual Conference of the Prognostics and Health Management Society, PHM, Philadelphia Center City, Philadelphia, PA, USA, 2018, pp. 24–27.

[59] G.B. Reid, T.E. Nygren, The subjective workload assessment technique: A scaling procedure for measuring mental workload, in: Advances in Psychology, Vol. 52, Elsevier, 1988, pp. 185–218.

[60] S.G. Hart, L.E. Staveland, Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, in: Advances in Psychology, Vol. 52, Elsevier, 1988, pp. 139–183.

[61] S.G. Hart, NASA-task load index (NASA-TLX); 20 years later, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 50, No. 9, Sage publications Sage CA, Los Angeles, CA, 2006, pp. 904–908.

[62] C. Lieber, Communications Between Air Traffic Controllers and Pilots During Simulated Arrivals: Relation of Closed Loop Communication Deviations to Loss of Separation (Ph.D. thesis), Arizona State University, 2020.

[63] J.C. Gorman, P.W. Foltz, P.A. Kiekel, M.J. Martin, N.J. Cooke, Evaluation of latent semantic analysis-based measures of team communications content, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 47, No. 3, Sage Publications Sage CA, Los Angeles, CA, 2003, pp. 424–428.

[64] E.E. Salas, S.M. Fiore, Team Cognition: Understanding the Factors that Drive Process and Performance, American Psychological Association, 2004.

[65] N.J. Cooke, J.C. Gorman, P.A. Kiekel, Communication as team-level cognitive processing, in: Macrocognition in Teams, CRC Press, 2017, pp. 51–64.

[66] N. Marwan, M. Carmen Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. 438 (5) (2007) 237–329, http://dx.doi.org/10.1016/j.physrep.2006.11.001, URL https://www.sciencedirect.com/science/article/pii/S0370157306004066.

[67] J.-P. Eckmann, S.O. Kamphorst, D. Ruelle, et al., Recurrence plots of dynamical systems, World Sci. Ser. Nonlinear Sci. Ser. A 16 (1995) 441–446.

[68] N. Marwan, M.C. Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. 438 (5–6) (2007) 237–329.

[69] M. Koebbe, G. Mayer-Kress, Use of recurrence plots in the analysis of time-series data, in: Santa Fe Institute Studies in the Sciences of Complexity-Proceedings Volume 12, Citeseer, 1992, p. 361.

[70] J.P. Zbilut, C.L. Webber Jr., Embeddings and delays as derived from quantification of recurrence plots, Phys. Lett. A 171 (3–4) (1992) 199–203.

[71] G.M. Mindlin, R. Gilmore, Topological analysis and synthesis of chaotic time series, Physica D 58 (1–4) (1992) 229–242.

[72] C.L. Webber Jr., J.P. Zbilut, Dynamical assessment of physiological systems and states using recurrence plot strategies, J. Appl. Physiol. 76 (2) (1994) 965–973.

[73] N. Marwan, N. Wessel, U. Meyerfeldt, A. Schirdewan, J. Kurths, Recurrence-plot-based measures of complexity and their application to heart-rate-variability data, Phys. Rev. E 66 (2) (2002) 026702.

[74] H. Kantz, Quantifying the closeness of fractal measures, Phys. Rev. E 49 (6) (1994) 5091.

[75] X. Li, X. Ying, M.C. Chuah, Grip: Graph-based interaction-aware trajectory prediction, in: 2019 IEEE Intelligent Transportation Systems Conference, ITSC, IEEE, 2019, pp. 3960–3966.

[76] A. Mohamed, K. Qian, M. Elhoseiny, C. Claudel, Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14424–14432.

[77] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[78] A. Pareja, G. Domeniconi, J. Chen, T. Ma, T. Suzumura, H. Kanezashi, T. Kaler, T. Schardl, C. Leiserson, Evolvegcn: Evolving graph convolutional networks for dynamic graphs, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 04, 2020, pp. 5363–5370.

[79] V. Vovk, A. Gammerman, C. Saunders, Machine-learning applications of algorithmic randomness, 1999.

[80] A.N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021, arXiv preprint arXiv:2107.07511.

[81] M. Sadinle, J. Lei, L. Wasserman, Least ambiguous set-valued classifiers with bounded error levels, J. Amer. Statist. Assoc. 114 (525) (2019) 223–234.

[82] A. Angelopoulos, S. Bates, J. Malik, M.I. Jordan, Uncertainty sets for image classifiers using conformal prediction, 2020, arXiv preprint arXiv:2009.14193.

[83] Y. Romano, M. Sesia, E. Candes, Classification with valid and adaptive coverage, Adv. Neural Inf. Process. Syst. 33 (2020) 3581–3591.

[84] A.N. Angelopoulos, S. Bates, A. Fisch, L. Lei, T. Schuster, Conformal risk control, 2022, arXiv preprint arXiv:2208.02814.

[85] H. Olsson, K. Kartasalo, N. Mulliqi, M. Capuccini, P. Ruusuvuori, H. Samaratunga, B. Delahunt, C. Lindskog, E.A. Janssen, A. Blilie, et al., Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction, Nat. Commun. 13 (1) (2022) 7761.

[86] V. Balasubramanian, S.-S. Ho, V. Vovk, Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications, Newnes, 2014.

[87] J. Alvarsson, S.A. McShane, U. Norinder, O. Spjuth, Predicting with confidence: using conformal prediction in drug discovery, J. Pharm. Sci. 110 (1) (2021) 42–49.
[88] C. Lu, A. Lemay, K. Chang, K. Höbel, J. Kalpathy-Cramer, Fair conformal predictors for applications in medical imaging, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 11, 2022, pp. 12008–12016.
[89] R. Luo, S. Zhao, J. Kuck, B. Ivanovic, S. Savarese, E. Schmerling, M. Pavone, Sample-efficient safety assurances using conformal prediction, in: Algorithmic Foundations of Robotics XV: Proceedings of the Fifteenth Workshop on the Algorithmic Foundations of Robotics, Springer, 2022, pp. 149–169.
[90] H. Papadopoulos, K. Proedrou, V. Vovk, A. Gammerman, Inductive confidence machines for regression, in: Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13, Springer, 2002, pp. 345–356.
[91] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Vol. 29, Springer, 2005.
[92] J. Lei, L. Wasserman, Distribution-free prediction bands for non-parametric regression, J. R. Stat. Soc. Ser. B Stat. Methodol. (2014) 71–96.
[93] A.A. Taha, A. Hanbury, Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool, BMC Med. Imaging 15 (1) (2015) 1–28.
[94] Y. Pang, H. Yao, J. Hu, Y. Liu, A recurrent neural network approach for aircraft trajectory prediction with weather features from sherlock, in: AIAA Aviation 2019 Forum, 2019, p. 3413.
[95] Y. Pang, X. Zhao, H. Yan, Y. Liu, Data-driven trajectory prediction with weather uncertainties: A Bayesian deep learning approach, Transp. Res. C 130 (2021) 103326.
[96] Y. Pang, X. Zhao, J. Hu, H. Yan, Y. Liu, Bayesian spatio-temporal graph transformer network (b-star) for multi-aircraft trajectory prediction, Knowl.-Based Syst. 249 (2022) 108998.