IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 March 2024, accepted 16 April 2024, date of publication 19 April 2024, date of current version 26 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3391427

RESEARCH ARTICLE

Air Traffic Controller Fatigue Detection Based on Facial and Vocal Features Using Long Short-Term Memory

ZHOUSHENG HUANG[1], WEIZHEN TANG[1], QIQI TIAN[2], TING HUANG[2], AND JINZE LI[2]
[1]CAAC Academy, Civil Aviation Flight University of China, Guanghan 618307, China
[2]College of Air Traffic Management, Civil Aviation Flight University of China, Guanghan 618307, China

Corresponding author: Zhousheng Huang (queensbarry@cafuc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62203451, and in part by Sichuan Science and Technology Project under Grant 2023JDRC0004.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Review Board of Civil Aviation Flight University of China.

ABSTRACT Air traffic controller fatigue has become a significant concern for flight safety. With the sharp rise in global air traffic, it is imperative to assess controller fatigue, as it directly impacts the safety and efficiency of air traffic control operations. Our study introduces a non-intrusive method to detect fatigue by analyzing the facial and vocal characteristics of air traffic controllers. Initially, we developed fast and accurate schemes for facial feature extraction, which allowed us to measure the ''percentage of eyelid closures'' and yawn frequency from video recordings. Subsequently, we extracted several vocal features from audio recordings, including average fundamental frequency, short-time average magnitude, short-time zero-crossing rate, harmonic-to-noise ratio, jitter, shimmer, loudness, and Mel-frequency cepstrum coefficient. We then created temporal sequences of these facial and vocal features to feed into a bi-directional long short-term memory gated recurrent unit network. This data, combined with the Stanford Sleepiness Scale, facilitated the identification and precise prediction of controller fatigue levels. Our experimental findings validate the effectiveness of the proposed detection method, which demonstrated a recognition accuracy rate of 95.12% on the test audio and video datasets.

INDEX TERMS Air traffic control, artificial intelligence, facial features, fatigue detection, long short-term memory, vocal features.

I. INTRODUCTION

The civil aviation industry is witnessing rapid growth, leading to a significant increase in the number of routes and aircraft operations. This growth has been accompanied by a rise in the complexity of airspace management, resulting in an increased workload for air traffic controllers (ATCs). Despite these challenges, the civil aviation sector continues to uphold strict safety standards. In response, the International Civil Aviation Organization (ICAO) has intensified its efforts to enhance the safety of air navigation systems. One of the initiatives introduced by the ICAO is the aviation system block upgrades framework [1], which aims to improve the safety and efficiency of air traffic management [2]. Recognizing that human factors play a crucial role in air transportation, fatigue among ATCs has become a focal point of attention. According to previous research [3], [4], fatigue is a major identifiable and preventable cause of accidents, accounting for 15% to 20% of all accidents. Addressing fatigue-related issues is essential for the sustainable development of the air transportation sector. Consequently, the ICAO has issued a series of guidelines and manuals on fatigue management [5], [6], [7], [8], [9]. The detection of ATC fatigue during operational duties is increasingly recognized as critical for maintaining safety and efficiency in air traffic management.

The associate editor coordinating the review of this manuscript and approving it for publication was Sunil Karamchandani.

Fatigue can lead to changes in human physiological and psychological processes, resulting in reduced human bodily functions that can cause diseases and accidents [10]. The pervasive nature of fatigue has made fully understanding it a considerable challenge for researchers worldwide [11]. ATCs are responsible for making critical decisions that can have an enormous impact on air traffic safety, decisions heavily influenced by their alertness, which is negatively affected by fatigue. Consequently, the need for more precise detection of ATC fatigue has become imperative.

Previous research has identified three main methods for fatigue detection:

(1) Physiological-indicator-based detection [12] utilizes methods such as electroencephalography (EEG), electrocardiography (ECG), electrooculography (EOG), functional near-infrared spectroscopy (fNIRS). Reference [13], and body temperature measurements. These methods employ sensors to collect signals, typically from the biceps or head of a subject.

(2) Questionnaire-based detection of personal fatigue levels. Reference [14] involves the use of common fatigue assessment scales, including the Karolinska sleepiness scale (KSS) [15], Stanford sleepiness scale (SSS) [16], Fatigue Scale-14. Reference [17], and National Aeronautics and Space Administration task load index [18]. Subjects complete a questionnaire, calculate the score after finishing it, and gauge their fatigue status based on the obtained score.

(3) Computer-vision based detection [19] involves capturing videos or images of the subject through computer-vision analysis, after which the status of the eyes, mouth, and body posture can be extracted. Indicators such as blinking and yawning frequency can then be analyzed to establish the level of fatigue of subject.

The above three methods are compared in **Table 1**.

With the continued development of machine learning (ML), numerous models have been developed. Both supervised and unsupervised learning models have been proposed and applied, including the support vector machine (SVM), convolutional neural network (CNN), recurrent neural network (RNN), dynamic fuzzy neural network (DFNN) [20], and long short-term memory (LSTM) models [21].

The aim of this study is to develop an ATC fatigue detection method by creating an LSTM-based fatigue detection model. Consequently, we explored LSTM models based on facial and vocal features to improve ATC fatigue detection accuracy. The following points summarize the main contributions of this work:

(1) *Extraction of facial and vocal features from video and audio*: We employed ML and voice analysis technology to simultaneously extract several facial and vocal features to describe the ATC fatigue states.

(2) *LSTM model for ATC fatigue detection*: We incorporated the aforementioned facial and vocal features into a traditional LSTM model to develop a non-intrusive model for ATC fatigue identification.

(3) *Matching the relationship between facial and vocal features and fatigue state*: We utilized the SSS to measure the ATC fatigue state and established the relationship between the facial features, voice features, and fatigue levels using the LSTM model.

In this study, by developing a non-invasive LSTM fatigue detection model based on facial and vocal features, we improved the detection accuracy. By using a pre-trained LSTM fatigue detection model, the real-time performance could be further enhanced.

II. RELATED WORKS

Current methods for detecting controller fatigue rely on physiological indicators, behavioral characteristics, or the voice characteristics of ATCs. There has also been a surge in proposals for fatigue detection methods based on ML by various researchers. Currently, fatigue detection research can be broadly categorized into two main types: subjective and objective detection methods. Subjective detection typically involves the use of fatigue scales and questionnaires to assess fatigue levels.

Conversely, objective fatigue detection uses facial features related to the eyes and mouth, based on computer-vision techniques. The simplest, most commonly used, and most effective metric is the percentage of eyelid closure (PERCLOS). Wierwille et al. [22] first employed PERCLOS to monitor fatigue in drivers. In 2010, Sommer and Golz [23] evaluated fatigue using PERCLOS, standard deviation of lateral position in a lane, EEG signals, and EOG signals. Additionally, they used the KSS to confirm the driver fatigue status, discovering a strong relationship between PERCLOS and KSS values. Their experiment revealed that PERCLOS was the most important indicator in fatigue detection. However, while the correlation between the EEG/EOG signal and fatigue was stronger than that of PERCLOS, the EEG/EOG evaluation was invasive.

With regard to the relevant features of the eye, Zhao et al. [24] proposed a CNN model that monitored the state of the eyes and mouth (EM-CNN) using region-of-interest (ROI) images to improve facial-fatigue recognition accuracy. In their study, when PERCLOS and mouth opening degree reached 0.25 and 0.5, respectively, the driver was considered to be in a state of fatigue. However, it must be noted that their experiments only classified fatigue and non-fatigue states without refining fatigue levels. Li et al. [25] incorporated a fatigue scale into the fatigue recognition process, along with an analysis of the driver's grip on the steering wheel. This combination was used to refine the assessment of the driver's fatigue level, yielding positive outcomes. Liang et al. [26] used an ES-DFNN model to detect controller fatigue, focusing primarily around the eye area.

The use of ML methods in fatigue detection has been gaining traction. In 2020, Zhao et al. [27] used the InceptionV3-LSTM model with multi-feature fusion to identify fatigue and sleepiness. They introduced an innovative approach using LSTM, showcasing the feasibility of utilizing LSTM for

TABLE 1. Three fatigue methods and their advantages and disadvantages.

Method	Advantages	Disadvantages
1	• Extremely high accuracy • Objective reflection of subject fatigue level	• Invasive • Subjectively affects the state of the subject, resulting in inaccurate results • Not suitable for use in ATC fatigue detection applications
2	• No contact between the device and the body • Simplicity • Only requires subjects to answer a few questions	• Highly subjective • Greatly affected by the emotions of the subjects
3	• Non-invasive • Simple equipment and low cost • Does not cause the subject to panic • Can be used in real-time fatigue detection and early warning applications to remind ATCs to avoid fatigue	• Greatly affected by pre-trained detection models • Results may differ from actual status

Note: Method 1 represents detection based on physiological indicators, Method 2 represents questionnaire-based detection of personnel fatigue levels, and Method 3 represents detection based on computer vision.

fatigue identification. In 2021, Chen et al. [28] used an LSTM model to detect driver fatigue based on facial key points and achieved highly accurate results. Their work underscored the concept that fatigue is a continuous behavior, suggesting that single-frame images might not yield the most precise evaluations. They highlighted the importance of considering fatigue's continuity and cumulative nature, issues addressable with LSTM. This challenge, centered on identifying the presence and degree of fatigue, essentially constitutes a binary or multi-classification problem. Akrout and Fakhfakh [29] also adopted an LSTM model, affirming its effectiveness in fatigue detection. Wang et al. [30] explored the use of the GLU-Oneformer model for fatigue detection, further contributing to the diverse approaches in this research area.

Kumar et al. [31] explored the use of image recognition and voice signals to detect COVID-19 infection in patients, showcasing the potential of ML techniques in medical diagnostics. This approach laid the groundwork for applying ML to identify fatigue levels through facial and vocal feature analysis. In 2023, Yu et al. [32] proposed a framework called RecMF for ATC mental fatigue (MF) recognition. RecMF employed an attention-enabled CNN-LSTM architecture that simultaneously captured time-series feature representations of EEG signals and eye movements. By using the RecMF framework, they found that increased levels of MF greatly reduced the reaction speed and accuracy of ATCs.

Milosevic [33] conducted research on bus and truck drivers and found that speech could reflect their level of fatigue. In 2014, Li et al. [34] proposed a fuzzy SVM model to detect driver fatigue based on nonlinear speech processing. They used the largest Lyapunov exponent, fractal dimension, and approximate entropy as indicators to judge the driver's level of fatigue. Their results confirmed the feasibility and effectiveness of using voice features for fatigue detection. Craye et al. [35] combined the PERCLOS and Mel-frequency cepstrum coefficient (MFCC) as inputs, before sending them to hidden Markov models and an SVM model to identify the state of driver fatigue. However, this method employed only a few facial and vocal features to identify fatigue; as a result, it recognized only whether the driver was fatigued or not. Using this method to detect controller fatigue would be insufficient, as it would be unable to identify the distinct fatigue states of controllers. Shen and Wei [36] proposed a high-precision feature extraction network to detect controller fatigue states using improved deep learning techniques. Gao et al. [37] developed a rapid and non-invasive method to assess the degree of fatigue based on connections between voice features and the level of fatigue as determined by the SSS. Their work highlighted those vocal characteristics could also reflect the level of fatigue, exhibiting strong correlations with the scale test values.

In 2022, Hu et al. [38] utilized CNN models to extract features for the eyes and mouth, added vocal features, and identified fatigue using the facial and vocal stacking (FV-stacking) method. The only speech feature used was the MFCC. The proposed model achieved 97% accuracy, with the best accuracy achieved by a single model being 92%. Similarly, the best accuracy realized by state-of-the-art detection methods was 88%. Consequently, driver fatigue detection and warning based on multi-information fusion represent a major development trend [39], as it is in the field of ATC fatigue detection.

TABLE 2. Advantages and disadvantages of different face recognition methods.

Method	Advantages	Disadvantages
MTCNN	● Specifically used for face recognition and labeling ● Few parameters and easy convergence ● High recognition accuracy ● No limit on image size or the number of faces	● Low push frame rate ● Too few negative samples can lead to misjudgment
Mask R-CNN	● Better accuracy than traditional the CNN method ● By introducing ROI alignment, the features and input can be aligned	● Classification boxes and prediction masks share evaluation functions, which can interfere with segmentation results
Dlib (with CNN)	● Easy to use ● Employed in many applications	● Low recognition accuracy ● Long labeling time

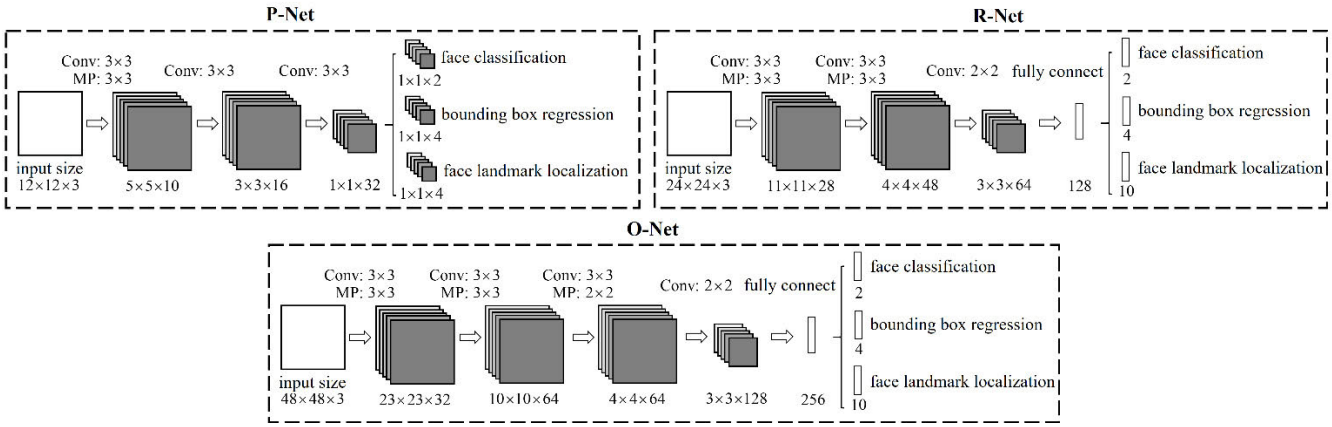


FIGURE 1. Structure of MTCNN model.

Our review of prior studies identified several limitations. Methods involving EEG, ECG, or EOG indicators necessitate the attachment of sensor electrodes to the subjects. The imposition of such electronic sensors on ATCs to monitor relevant indicators is impractical, as it would encumber them and potentially cause discomfort, thereby influencing the outcomes of the experiment. In light of these challenges, we explored an alternative approach that involves extracting facial and vocal features from video and audio recordings. This approach allows a pre-trained ML model to identify the pertinent descriptors without causing inconvenience to the ATCs.

Fatigue in the human body is not only manifested in the face or speech, but these aspects can often reflect a person's level of fatigue simultaneously. However, most studies have only explored the relationships between facial or vocal features and fatigue independently, with only a few studies having combined the two. We also found that using ML-based methods—notably LSTM-based methods with forget gates to identify fatigue—has yielded satisfactory results.

The following section describes the proposed model, which includes facial and vocal features combined with a fatigue assessment scale. This model learns from existing data using the LSTM model to more accurately identify the level of fatigue, instead of simply recognizing the two states of wakefulness and fatigue.

III. METHODOLOGY

A. FACE AND FACIAL KEY-POINT DETECTION

Face and facial key-point detection are among the most critical steps in the proposed method. However, recognizing faces and marking key points across different controller postures and states can be challenging, with subsequent eye and mouth feature extraction based on the identification of these key points. To identify the key facial points, the face area must first be quickly detected. With the development of ML, various face detection methods have emerged, including the multi-task CNN (MTCNN) [40], mask R-CNN [41], and the *dlib* library (with CNN) [42], [43], [44]. The advantages and disadvantages of each of these three-face recognition and key-point labeling schemes are summarized in **Table 2**. After comparing the three facial recognition methods, we chose the MTCNN model to determine the face area of collected images. The structure of the MTCNN model [40] is shown in **Fig. 1**.

The MTCNN model comprises three sequential CNNs—namely, the proposal network (P-Net), refined network (R-Net), and output network (O-Net). The MTCNN method is a rough-to-fine process, with the functions of each network being as follows:

P-Net: A shallow CNN that quickly screens out potential candidate boxes for faces. Its output comprises three parts: (1) binary classification results, representing whether a face

exists; (2) the position of the detected face frame; and (3) the position of the key points within the detected face frame.

R-Net: A more complex CNN used to remove the proposals that do not contain faces, primarily to correct the results of the P-Net and eliminate errors.

O-Net: This CNN is similar to the P-Net but has larger input dimensions and depth, and the ability to output more accurate results.

After a face image is obtained by the MTCNN model, landmarks can be extracted from it. Common face feature point recognition frameworks include the *dlib* library (with default face detector) and the MediaPipe framework [45], with their basic conditions and differences as listed in **Table 3**.

TABLE 3. Comparison of DLIB library (with default face detector) and mediapipe results.

Toolkits	Number of landmarks	Advantages
Dlib (with default face detector)	68	• Relatively simple to use
MediaPipe	468	• More stable than *dlib* • High FPS revolution [46] • Three-dimensional coordinates can be estimated

The two toolkits mentioned above were compared. Because MediaPipe can recognize more landmarks and has a higher FPS, it could achieve better performance in real-time detection; consequently, we chose MediaPipe for landmark marking. We obtained the i^{th} landmark coordinates (x_i, y_i), each landmark being as shown in **Fig. 2**.

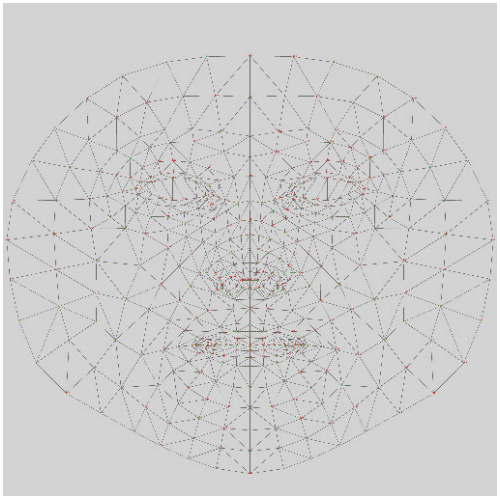
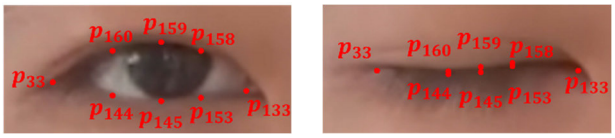


FIGURE 2. Visualization of the 468 facial landmark coordinates using MediaPipe.

B. FACIAL FEATURES

1) EYE FEATURES

Eye blink detection plays a crucial role in monitoring operator fatigue, including that of ATCs. To pinpoint the blinking characteristic, the method involves utilizing numbered landmarks, which are highlighted in conjunction with facial key-point marking techniques (as shown in **Fig. 3**). These landmarks define the feature points around the eyes. In [47], only six key points were used to describe the height and width of the eyes. To more accurately describe their height, we included two additional key points.



(a) (b)

FIGURE 3. Opened (a) and closed (b) left eye with landmarks (p_i).

As shown in **Fig. 3**, when MediaPipe is used to analyze the key points of the human eye, the description of the eye can be divided into inner and outer eye circles. To describe the eye features more accurately, we used the inner eye circle for analysis. In the proposed model, the eye aspect ratio (EAR) [47] between the height and width of the eyes can be expressed as follows:

$$\mathrm{EAR_L} = \frac{\|p_{160} - p_{144}\| + \|p_{158} - p_{153}\| + 2\,\|p_{159} - p_{145}\|}{4\,\|p_{33} - p_{133}\|}, \tag{1}$$

$$\mathrm{EAR_R} = \frac{\|p_{386} - p_{374}\| + \|p_{385} - p_{380}\| + 2\,\|p_{387} - p_{373}\|}{4\,\|p_{362} - p_{263}\|}, \tag{2}$$

where $\mathrm{EAR_L}$ denotes the EAR of the left eye, $\mathrm{EAR_R}$ denotes the EAR of the right eye, and p_i denotes the two-dimensional (2D) landmark locations defined in **Fig. 3**. As is evident, the EAR is constant when the eye is open and close to zero when closed. The EAR of the open eye varies slightly among individuals but is fully invariant to uniform scaling of the image and in-plane rotation of the face. With increasing fatigue, the eyes may remain closed for longer time periods; consequently, the EAR decreases.

Previous studies [23] have shown that the variation of the PERCLOS over time is a good fatigue indicator; the PERCLOS is the percentage of eyelid closure over the pupil over time and reflects slow eyelid closures (''droops'') rather than blinks. In real-time feature recognition based on video signals, this value can be obtained as follows:

$$\mathrm{PERCLOS} = \frac{m}{M} \times 100\%, \tag{3}$$

where m denotes the total time the eyes are closed during the statistical time period and M denotes the total time counted.

Three methods exist for the calculation of the PERCLOS—namely, the EM, P70, and P80 threshold values. These methods are compared in **Table 4**.

By using the EAR method, the ratio of the eyelid to the eyeball can be calculated. As shown in **Fig. 4**, the average value of the EAR for both eyes is approximately 0.25. If EAR

TABLE 4. Meaning of EM, P70, and P80 threshold values.

Criterion	Paraphrase
EM	The eyelid covers the eyeball more than 50% of the time
P70	The eyelid covers the eyeball more than 70% of the time
P80*	The eyelid covers the eyeball more than 80% of the time

* The PERCLOS values in this paper were obtained using P80 as the criterion when not specifically mentioned.

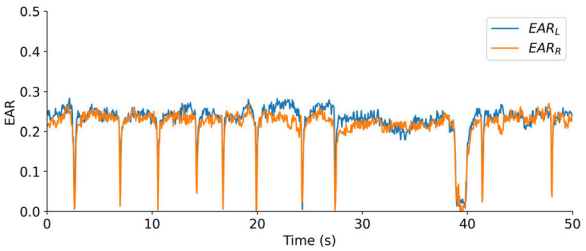


FIGURE 4. Example of EAR variation over time.

$= 0.25$ is taken to be the standard for eye opening, then for the P80 evaluation method, when EAR < 0.05, the time can be calculated to obtain the value of m.

2) MOUTH FEATURES

In addition to eye features, we considered mouth features, which can be used to identify whether a person is yawning. By using MediaPipe, we were able to obtain the mouth landmarks, as shown in **Fig. 5**.

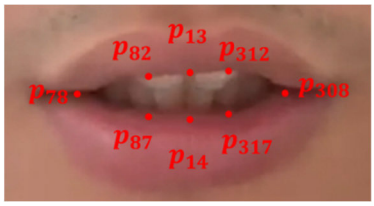


FIGURE 5. Landmarks of the mouth.

MediaPipe employs numerous landmarks for delineating mouth features, akin to those used for the eyes. In our analysis, we focus on the innermost point for our calculations. Consequently, we can define the mouth aspect ratio (MAR) using the following formula:

$$\mathrm{MAR} = \frac{\|p_{82} - p_{87}\| + \|p_{312} - p_{317}\| + 2\,\|p_{13} - p_{14}\|}{4\,\|p_{78} - p_{308}\|}, \tag{4}$$

where p_i denotes the 2D landmark locations, as defined in **Fig. 5**.

As shown in **Fig. 6**, the MAR is close to zero when the subject is silent and approximately 0.2 when the subject is speaking normally. A yawning behavior is indicated when the MAR surpasses 0.4. By tracking the MAR fluctuations in real-time from the source videos, it is possible to detect yawning. In this study, an MAR value exceeding 0.4 that persists for at least 2 s is interpreted as a yawning event, which is a recognized indicator of fatigue. Utilizing this criterion, we can quantify the total duration of yawning episodes.

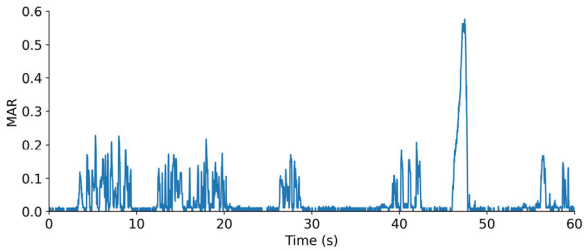


FIGURE 6. Example of MAR.

C. VOICE FEATURES

1) FUNDAMENTAL FREQUENCY

The fundamental frequency (*F0*) of a speech signal refers to the approximate frequency of the periodic structure of speech signals, and it is one of the most commonly used indicators to describe the sound. Slight differences in the *F0* in different states may exist. Consequently, we used the probabilistic YIN [40] algorithm to estimate the *F0* of a segment of voice.

2) SHORT-TIME AVERAGE MAGNITUDE

The short-time average amplitude can be used to analyze the energy distribution of a speech signal. By calculating the average amplitude within a period of time, the overall energy level of the sound within that period can be obtained. This information helps us understand the energy characteristics of the speech signal and to compare changes in the sound intensity over time.

Here, we use M_n to denote the short-time average magnitude of the speech signal in the n^{th} frame, which can be expressed as follows:

$$M_n = \frac{1}{N}\sum_{m=0}^{N-1} |x_n(m)|, \tag{5}$$

where $x_n(m)$ denotes the n^{th} frame of the speech signal, and N denotes the total number of frames of speech.

3) SHORT-TIME ZERO-CROSSING RATE

A specific relationship exists between the zero-crossing rate and the clarity and noise components of a sound. A higher zero-crossing rate generally indicates higher clarity of sound, while a lower zero-crossing rate suggests noise or muffled sounds; moreover, zero-crossing rates are possibly connected to fatigue.

For the speech signal $x_n(m)$ of the n^{th} frame, the short-term zero-crossing rate can be expressed as follows:

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1} |\mathrm{sgn}(x_n(m)) - \mathrm{sgn}(x_n(m-1))|, \tag{6}$$

where sgn(x) denotes a symbolic function:

$$\mathrm{sgn}(x) = \begin{cases} 1 & , x \geq 0 \\ -1 & , x < 0. \end{cases} \tag{7}$$

4) HARMONIC-TO-NOISE RATIO (HNR)

The purpose of the harmonic-to-noise ratio (HNR) is to quantify the relative proportion of harmonic and noise components in a speech signal, with the aim of assessing speech clarity and the level of noise interference. The HNR can quantitatively capture changes between the harmonic and noise components in sound characteristics, providing an objective indicator to quantify the impact of fatigue on them. For example, fatigue can lead to a decrease in speech quality, and the HNR can be used to evaluate the proportional change of harmonic and noise components in speech signals, providing a quantitative index to measure speech quality. Notably, the HNR value may decline under fatigue.

The HNR is a logarithmic measure of the energy ratio associated with the harmonic and noise components and can be defined as follows:

$$\mathrm{HNR} = 10 \lg \frac{\int_w |H(w)|^2}{\int_w |N(w)|^2}, \tag{8}$$

where $H(m)$ denotes the harmonic component, and $N(m)$ denotes the noise component.

5) JITTER

Jitter is a measure used to describe the irregularity and instability of changes in the *F0* [48]. Fatigue can affect the function of the vocal cords and speech system, leading to abnormal changes in *F0*. By calculating the jitter, the fatigue state of the sound can be characterized. Jitter [49] can be defined as follows:

$$j = \frac{\frac{1}{n-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\frac{1}{N} \sum_{i=1}^{N} T_i}, \tag{9}$$

where J denotes the jitter, T_i denotes the extracted period lengths, and N denotes the number of extracted periods.

6) SHIMMER

Shimmer is an index used to quantify irregular changes and the instability of sound amplitude. By calculating the shimmer value, the sound amplitude changes can be assessed. A higher shimmer value generally indicates that the sound is more unstable, whereas a lower shimmer [49] value indicates that the sound is more stable. Shimmer can be defined as follows:

$$S = \frac{\frac{1}{n-1} \sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\frac{1}{N} \sum_{i=1}^{N} A_i}, \tag{10}$$

where S denotes the shimmer, A_i denotes the extracted peak-to-peak amplitude data, and N denotes the number of extracted *F0* periods.

7) LOUDNESS

Loudness is an indicator used to quantify sound intensity or volume. Correspondingly, fatigue can cause changes in sound intensity. By observing changes in sound loudness, the characteristics of sound under fatigue can be captured. Loudness can be defined as follows:

$$L = 10 \lg \frac{I}{I_0}, \tag{11}$$

where L denotes the loudness of the sound (in decibels), I denotes the sound pressure level, and I_0 denotes the reference sound pressure level (usually 20μPa).

8) MFCC

The MFCC is a cepstral parameter extracted from the Mel scale frequency domain, which can be used to describe nonlinear characteristics related to the frequencies of the human ear [50]. Generally, in speech recognition tasks, the MFCC is considered a feature vector describing the sound signal of each frame—that is, the MFCC is a feature processing index that simulates the way in which the human auditory system processes sound and can be used to distinguish between different speakers and speech states [51], [52], [53].

MFCC analysis focuses on the auditory characteristics of the human ear, noting that the level of sound heard by the human ear is not linearly proportional to its frequency. Incorporating this frequency enables the features to match more closely the sounds humans hear. The formula for converting from the actual voice frequency to the Mel-frequency is as follows:

$$f_{\mathrm{mel}} = 2595 \lg(1 + \frac{f}{700}). \tag{12}$$

Fig. 7 depicts a flow chart of the process to obtain the MFCC.

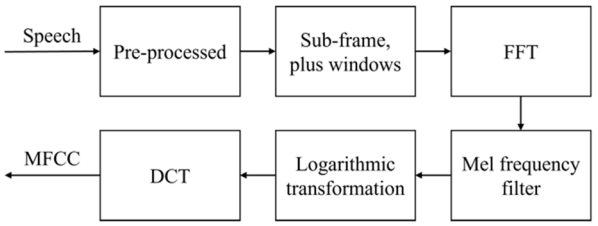


FIGURE 7. Flow chart of the MFCC extraction process.

To extract the MFCC, the speech signal needs to be pre-processed first to ensure that high-frequency components are not distorted. The skipped signal can then be smoothed by the sub-frame plus windows [54] to decompose the sound signal into a series of overlapping time windows, with a window function (Hamming window) being applied to each

time window. A fast Fourier transform can be performed to obtain the speech signal spectrum. The specific spectrum, Mel filter bank, and spectrum envelope can be extracted using a Mel-frequency filter. Subsequently, a logarithmic transformation can be applied to the signal to obtain the logarithmic spectrum. Finally, the MFCC feature can be acquired by applying the discrete cosine transform.

The MFCC can access relevant dimensions depending on the requirements. In this study, based on a previous study [37], we considered using 12 MFCCs (1)–(12) as a part of the speech features.

D. FATIGUE DETECTION NETWORK BASED ON LSTM

The LSTM [55] is a time-recurrent neural network designed to solve the long-term dependence problem of general RNNs. The LSTM network is suitable for processing and predicting important events with very long intervals and delays in time-series. While it possesses a chain structure similar to that of the RNN, as shown in **Fig. 8**, it differs from an RNN cell; a detailed introduction to the LSTM model can be found in Appendix.

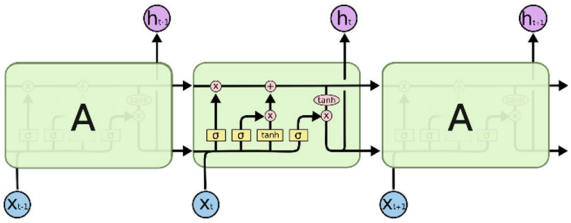


FIGURE 8. General LSTM structure.

By using facial and vocal feature extraction, we set the feature vector to 21. Consequently, the controller fatigue detection feature vector can be expressed as follows:

$$\mathbf{F}_n = \left[f_{n,1} \cdots f_{n,21} \right], \tag{13}$$

where the corresponding relationship of each value in the eigenvector is shown in **Table 4**, and $\mathbf{F}_n$ is used as an input into the LSTM.

Using the steps presented in **Fig. 10**, we first obtained several video samples; here, we use one video as an example. Next, we divided the video sample into several video segments of equal length, extracted facial and vocal features using the proposed method, formed a feature matrix of the video segments, input it into the LSTM network for training, performed multi-classification using the SoftMax function, and finally determined the fatigue level. The feature matrix can be expressed as follows:

$$\mathbf{F} = \begin{bmatrix} f_{11} & \cdots & f_{1m} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nm} \end{bmatrix}, \tag{14}$$

where the vertical direction represents the length of the time-series, which is equal to the number of video clips, and the horizontal direction represents the number of features—that is, $m = 21$ as per **Table 5**.

In this process, video segments are used as inputs, so the segment lengths need to be carefully considered. If the length of the segment is too long, the recognition may be insufficient, resulting in low accuracy in practical applications. If the length of the segment is too short, the feature value error may be too large, affecting the operation of the LSTM model. To capture key yawning information in video segments, and consistent with existing research [56], the average duration of human yawning is assumed to be 6.5 s. Therefore, a video segment should be at least 6.5 s in duration to provide sufficiently complete information. Moreover, the normal speaking speed of a person is between 160 and 180 Chinese characters per minute.

Consequently, to achieve better results, we considered setting each video segment to 20 s. For the LSTM network, to obtain more samples and ensure the continuity of samples, we utilized the sliding window concept, as shown in **Fig. 10,** to acquire video clips. Thus, based on a comprehensive analysis of **Fig. 9**, and combined with the segmentation scheme shown in **Fig. 10**, **Table 4**, and (13), if the duration of a video sample is 5 min, then the dimensions of its feature matrix should be 280×21.

In the proposed fatigue recognition model, for the LSTM model, the degree of fatigue evident in a video sample should be marked. Therefore, we chose the SSS to delineate the fatigue values, the corresponding SSS values of which are shown in **Table 6**.

When filling in the SSS, the scale rating X will not appear. Additionally, two marks for the video clips were made: one based on the scale rating of the subject according to the SSS, and the other on the scale rating to categorize the video segments into "*awake*" or "*sleepy*" depending on the amount of work completed by the controller in the real world. This corresponds to the scale ratings and relationships shown in **Table 7**.

IV. EXPERIMENTAL RESULTS

A. PARTICIPANTS

In this study, we recruited 40 active ATCs. The participants included both men and women, all aged between 25 and 35. Moreover, to ensure the accuracy of the experiment, they had all been engaged in regulatory work for at least 2 years (inclusive). Their basic characteristics are shown in **Table 8**.

B. EXPERIMENTAL DATASET

1) SELF-BUILT DATASET

During this study, we deprived all participants of sleep for 48 h without limiting their participation in other activities to obtain a diverse sample dataset. During the experiment, we asked them to fill in the SSS questionnaire every hour to collect their feedback; after filling out the questionnaire, we let them work in a simulated supervisory position for 5 min to collect their facial and vocal data.

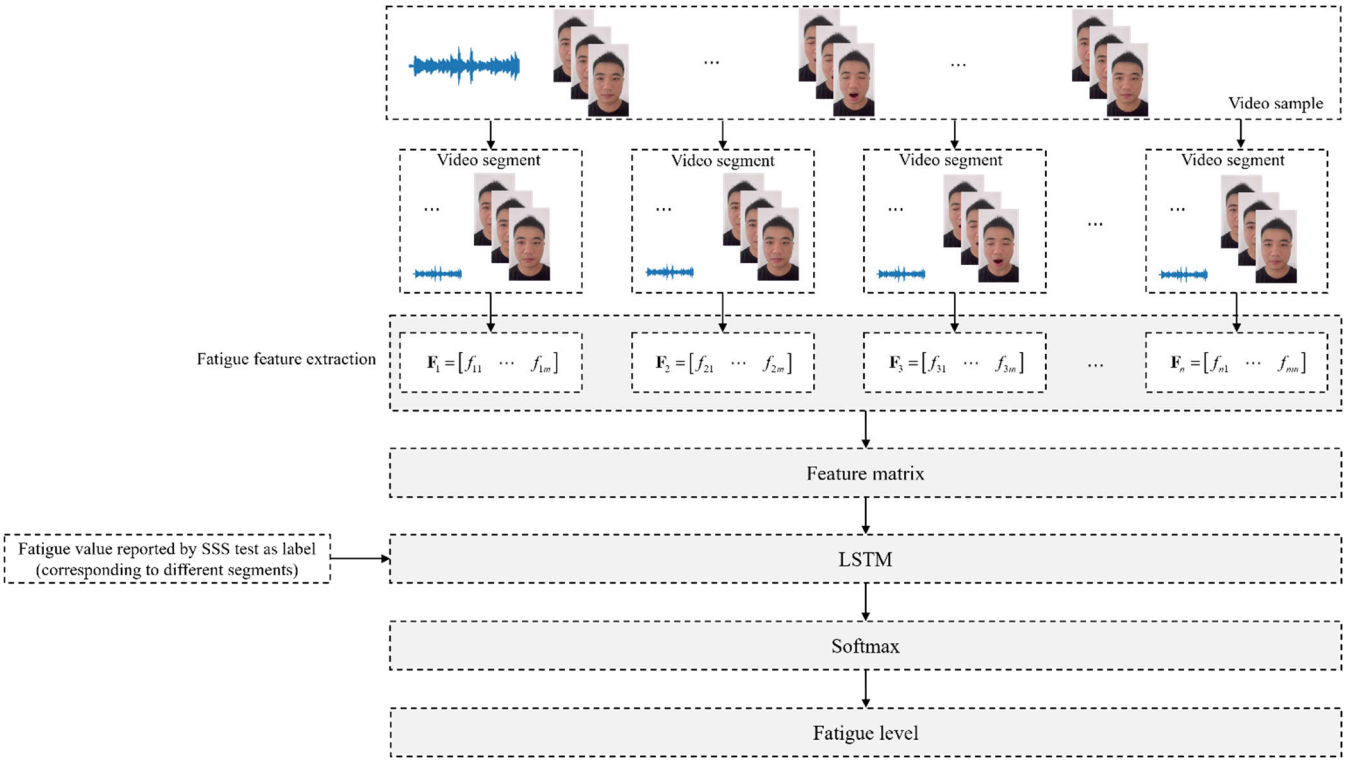


FIGURE 9. Flow of the proposed fatigue detection method for a sample.

TABLE 5. Relationship between each value in the feature vector and facial or vocal features.

Value	Feature	Description
$f_{n,1}$	PERCLOS	Obtained by analyzing the duration of eye closure in video segments
$f_{n,2}$	Number of yawns	Number of yawns in video clips identified based on the MAR
$f_{n,3}$	Average *F0*	Average *F0* in video segments with audio
$f_{n,4}$	M_n	Short-time average amplitude in video segments with audio
$f_{n,5}$	Z_n	Short-time zero-crossing rate in video segments with audio
$f_{n,6}$	HNR	HNR in video segments with audio
$f_{n,7}$	Jitter	Jitter in video segments with audio
$f_{n,8}$	Shimmer	Shimmer in video segments with audio
$f_{n,9}$	L	Loudness in video segments with audio
$f_{n,10}\cdots f_{n,21}$	MFCC	1–12 dimensions MFCC in video segments with audio

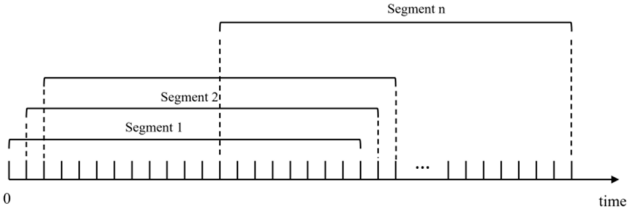


FIGURE 10. Schematic of video segment acquisition from a video sample.

We then trimmed the collected video and audio data based on the proposed 20-s standard, obtaining a total of 1,920 samples of audio and video data with corresponding fatigue values from 1 to 7. Some of the images in the dataset are shown in **Fig. 11**.

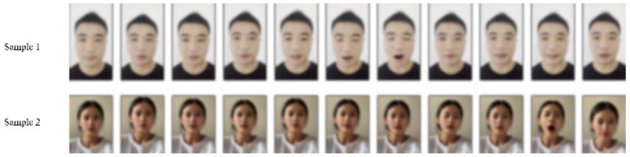


FIGURE 11. Sample images in dataset.

TABLE 6. SSS.

Degree of sleepiness	Scale rating
Feeling active and vital; alert; wide awake	1
Functioning at a high level, but not at peak; able to concentrate	2
Relaxed; awake; not at full alertness; responsive	3
A little foggy; not at peak; let down	4
Fogginess; beginning to lose interest in remaining awake; slowed down	5
Sleepiness; prefer to be lying down; fighting sleep; woozy	6
Almost in reverie; sleep onset soon; lost struggle to remain awake	7
Asleep	X

TABLE 7. Relationship between scale ratings and wakefulness and sleepiness.

Scale rating	Wakefulness or sleepiness
1–3	Wakefulness
4–7	Sleepiness

TABLE 8. Participant characteristics.

Characteristic	Value
Number of participants	40
Percentage of male participants*	75%
Percentage of female participants	25%
Age (years)	29.2±3.3
Working experience	5±2.4

All subjects have supported and agreed to the experiment.
* China has significantly more male ATCs than female ATCs, so the participants were mostly male.

2) UNIVERSITY OF TEXAS AT ARLINGTON REAL-LIFE DROWSINESS DATASET (UTA-RLDD)

The University of Texas at Arlington real-life drowsiness dataset (UTA-RLDD) [57] was created for multi-stage drowsiness detection, targeting not only extreme and easily visible cases, but also cases when subtle micro-expressions were the discriminative factors. The dataset contains 60 frontal videos of different people performing a simple task (reading or watching something on a computer), each recorded for 10 min. Three categories of videos are provided in the UTA-RLDD—namely, awake, low-vigilance, and drowsy—similar to the dataset we built ourselves. Both datasets involve situations in which the subjects face a computer screen and are related to control work. It is worth noting that the videos in this dataset do not contain sound, so there are certain differences from our self-built dataset. An example of the UTA-RLDD dataset is shown in **Fig. 12**.

C. EXPERIMENTAL ENVIRONMENT

The experimental platform used in this study included the following: (1) a Windows 11 operating system running on an Intel Core i7-9700K CPU, with an NVIDIA GeForce GTX 1050 18 GB independent graphics card, and 16 GB of memory; (2) a 1080p resolution camera and microphone with a 768 kbps bitrate; and (3) PyTorch, which was used to build the ATC fatigue detection network.

FIGURE 12. Sample images in the UTA-RLDD.

D. STATISTICAL ANALYSIS

First, we analyzed the scale ratings, the results of which are shown in **Fig. 14.**

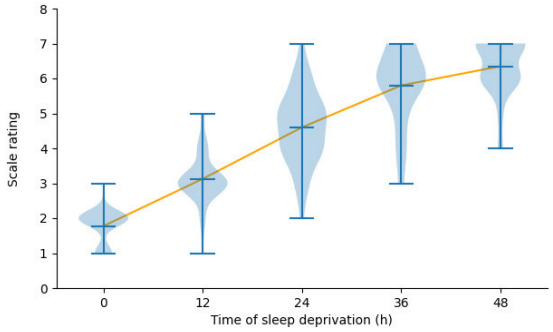


FIGURE 13. SSS scores under different sleep deprivation times.

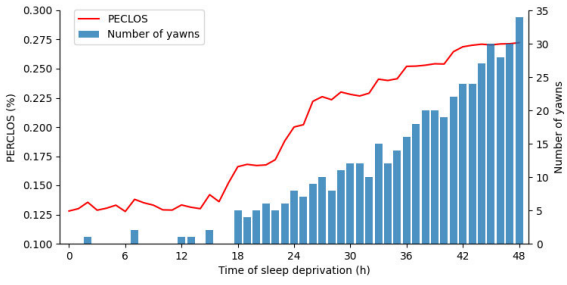


FIGURE 14. PERCLOS and number of yawns.

In **Fig. 13**, the blue shaded part represents the occupied area corresponding to the scale score. It demonstrates that the SSS better represents degree of fatigue of individuals; moreover, when the sample size is large, it satisfactorily explains the changing fatigue trends. Clearly, as the experiment progresses, the average score value of the scale, illustrated by the orange line, increases considerably, and the average

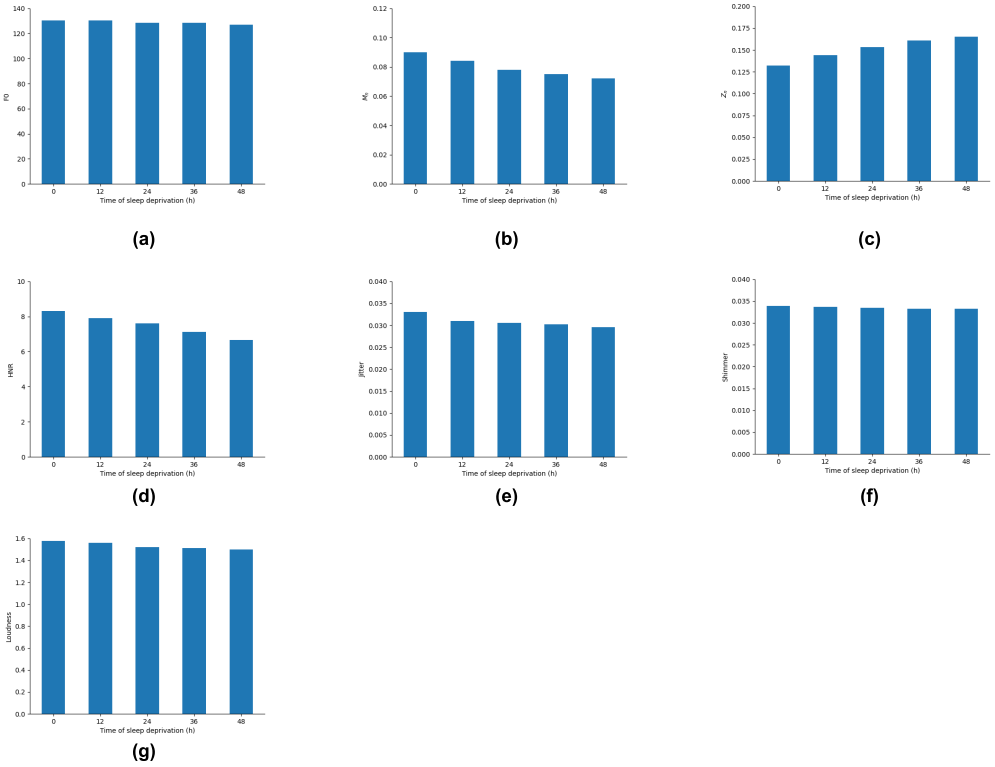


FIGURE 15. Voice features after 48 h of sleep deprivation: (a) Average F0; (b) (M_n); (c) (Z_n); (d) HNR; (e) Jitter; (f) Shimmer; (g) Loudness.

fatigue level of all participants increases. Further, the samples collected occupy each scoring interval.

With reference to previous research, we then analyzed the PERCLOS and number of yawns per hour, the statistical results of which are shown in **Fig. 14**.

Figs. 13 and **14** together reveal that with the increasing average fatigue levels during the experiment, the average PERCLOS value and number of hourly yawns generally trend upward.

Here, for two sequences of equal length, the Pearson correlation coefficient can be used to calculate the degree of correlation between two variables and can be expressed as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{15}$$

where X and Y denote two sequences.

The correlation coefficient between the PERCLOS and yawn values is 93.55% based on (15). Consequently, a strong correlation exists between the PERCLOS and yawn count. Moreover, the hourly PERCLOS and mean scale scores and the hourly number of yawns and mean scale scores can be calculated using the Pearson correlation coefficients, the correlations being 96.86% and 90.04%, respectively. The above data prove that a strong correlation exists between the three aforementioned quantities and that using the PERCLOS and yawn count as inputs to the fatigue rating network is reasonable.

For the vocal features, variations are shown in **Fig. 15**. It is evident that the sound feature values all exhibit regular changes as the experiment proceeds. As speech also exhibits regularized variation, adding speech features to the LSTM network can improve the prediction accuracy of this network.

E. ANALYSIS OF RESULTS

To clearly describe the effectiveness of the proposed algorithm, ablation and comparison experiments were conducted on fatigue detection, and predictions of different networks were compared, the results of which are discussed below.

1) RESULTS OF USING OUR NETWORK ON THE SELF-BUILT DATASET

In this study, we used accuracy, precision, recall, and F1-score values to assess the model classification effectiveness. These quantities can be expressed as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{16}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{18}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{19}$$

where the meanings of TP, FP, TN, and FN are shown in **Table 9**.

TABLE 9. Meanings of TP, FP, FN, AND TN.

Abbreviation	Full name	Meaning
TP	True positive	Positive samples and positive model prediction
FP	False positive	Negative samples and positive model prediction
TN	True negative	Positive samples and negative model prediction
FN	False negative	Negative samples and negative model prediction

The network outlined in **Fig. 9** was utilized, allocating 70% of the amassed 1,920 data samples for training and the remaining 30% for testing. Initially, we categorized the scale outcomes into two groups as per the criteria set forth in **Table 7**. These categorized results were then fed into the proposed network for predictive analysis. The outcomes of this analysis are presented in **Table 10**, with the corresponding confusion matrix depicted in **Fig. 16**.

TABLE 10. Model evaluation metrics in binary classification on self-built test dataset.

Status	Accuracy (%)	Precision (%)	Recall (%)	F1-score
Wakefulness	97.23	96.70	97.24	0.9697
Sleepiness	98.48	98.73	98.48	0.9861
Overall status	98.09	98.09	98.09	0.9809

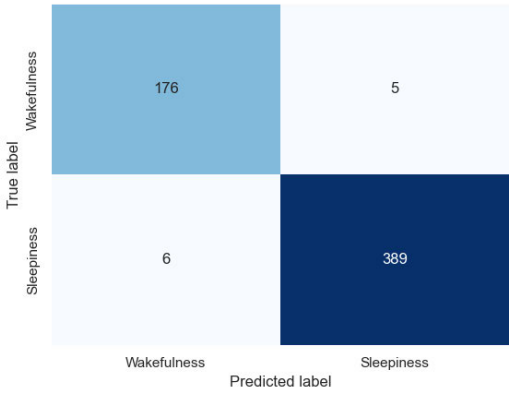


FIGURE 16. Confusion matrix under binary classification conditions.

Based on **Table 9** and **Fig. 16**, using the proposed ATC fatigue network to predict whether controllers were awake or fatigued proved to be effective—that is, this prediction network was accurate when conducting binary classification. We then used it for multi-classification prediction, the results of which are presented in **Table 11** and **Fig. 17**.

Fig. 18 demonstrates how the scale rating values evaluate the indicators in the model.

A considerable decrease in accuracy occurs when using the proposed model for specific fatigue prediction compared to predicting only fatigue or wakefulness. However, the

TABLE 11. Model evaluation metrics in multi-classification on self-built test dataset.

Scale rating	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
1	96.77	90.91	96.77	0.9375
2	86.30	94.03	86.30	0.9000
3	93.51	86.75	93.51	0.9000
4	91.00	91.92	91.00	0.9146
5	94.68	91.75	94.68	0.9319
6	87.61	93.40	87.61	0.9041
7	96.59	93.41	96.59	0.9497
Overall rating	91.84	91.93	91.84	0.9182

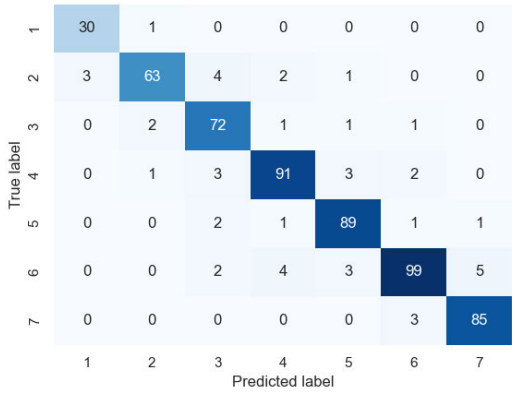


FIGURE 17. Confusion matrix under multi-classification conditions.

overall accuracy is maintained at 91.84%, which is effective for fine-grained assessment of ATC fatigue. Moreover, high accuracy, recall, and F1-score values were obtained, indicating the superior performance of the model in all respects. As is evident from **Fig. 18**, for the prediction of the four values of 1, 4, 5, and 7 scale ratings, the model performs well on each index, with these values being concentrated in the awake state, general fatigue state, and extreme fatigue state. The predictions corresponding to the 2 and 6 scale ratings are less accurate and precise in terms of prediction, as these states are transition states.

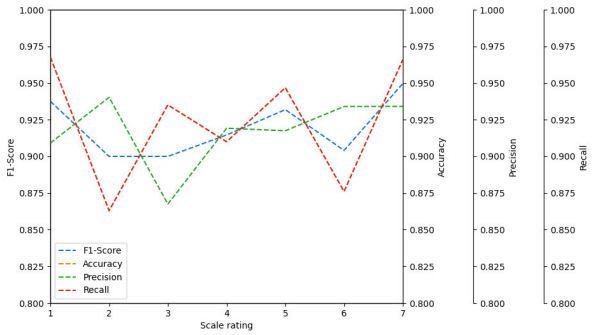


FIGURE 18. Metrics in the proposed model (accuracy is same as recall).

2) USE OF VARIANT LSTM NETWORKS

We employed the traditional LSTM network for ATC fatigue prediction, given its superior ability to process temporal data compared to other methodologies. This approach

TABLE 12. Model evaluation metrics in multi-classification on self-built test dataset.

Network	Accuracy (%)	Advantages over traditional LSTM
Traditional LSTM (Previously used)	91.84	—
Bi-LSTM	92.97	The network a bi-directional LSTM network, giving it the ability to extract information from both above and below, enabling more comprehensive perception of time-series type data
LSTM+GRU [58]	92.88	The network synthesizes forget and input gates into a single update gate, and likewise mixes cell states and hidden states
Bi-LSTM+GRU [59], [60]	**93.83**	The advantages of Bi-LSTM and GRU models are combined

yielded a notably high accuracy. Motivated by these results, we explored a variant of the LSTM network to assess the potential for further enhancing the prediction accuracy. The accuracy achieved with this modified LSTM method is detailed in **Table 12**.

As is evident from **Table 12**, the use of a variant LSTM network improves the prediction accuracy. Using a combination of the Bi-LSTM network and gated recurrent unit (GRU) results in an accuracy improvement of 1.49% compared to the traditional LSTM network. The variant networks tend to focus more on the processing of the time-series data and the updating of the cell states, usually improving the performance of the model. Consequently, the use of variant LSTM networks to identify and predict the exact fatigue level of controllers is appropriate; in this study, the combined Bi-LSTM and GRU model was selected for further optimization.

3) HYPERPARAMETER OPTIMIZATION OF BI-LSTM AND GRU MODELS

After comparing four LSTM models and variants, the accuracy obtained by the Bi-LSTM+GRU model was found to be the most accurate. In previous experiments, the parameter values used were all default hyperparameter values, which achieved an accuracy of approximately 94%. The training and test datasets in the previous experiments had only one time division, meaning the test dataset was fixed.

Assessing the performance of one model on one test dataset does not provide the best indication of how the model will perform over a wide variety of test data.

To further improve model performance, we implemented K-fold cross-validation. References [61] and [62] in combination with grid search [63], [64] to hyperparameter tuning [65]. The outcomes of this process are shown in **Table 13**. Additionally, we compared the results achieved with the Bi-LSTM+GRU network before and after applying the optimized combination of hyperparameters, with this comparison detailed in **Table 14**.

As is evident from **Table 14**, the accuracy of the optimized model reaches 95.12%, which is 1.29% higher than that of the model before optimization.

Although the proposed LSTM model can predict ATC fatigue levels well, it cannot provide prediction confidence. To solve this problem and to determine and verify the credibility of the proposed model in the case of small samples, the Monte Carlo simulation method [66], [67], [68] was used to conduct confidence testing. We conducted 1,000 Monte Carlo simulation trials, incorporating random noise into each. The model's accuracy, optimized to 95.12% through hyperparameter tuning as shown in **Table 14**, surpassed the 95% confidence threshold, affirming the model's reliability for this specific dataset. The optimal hyperparameter set was identified using K-fold cross-validation, which involves segmenting the dataset into multiple parts and iteratively training and evaluating the model across these segments. This approach confirmed the model's robust performance across different data subsets. The combination of these rigorous methodologies and the positive outcomes from the Monte Carlo simulations underlines the model's credibility and efficacy.

TABLE 13. Optimal combination of hyperparameters using K-fold cross-validation and grid search.

Hyperparameter	Range	Step length	Optimized value
Neurons	[1, 50]	1	16
Learning_rate	[0, 1]	0.001	0.005
Dropout	[0, 1]	0.01	0.1
Optimizer	ADAM, SGD	—	ADAM
Recurrent_dropout_rate	[0, 1]	0.01	0.0

TABLE 14. Comparison of network accuracy before and after optimization.

Network	Accuracy (%)
Bi-LSTM+GRU (unoptimized)	93.83
Bi-LSTM+GRU (optimized)	**95.12**

4) COMPARISON WITH DIFFERENT NETWORKS ON SELF-BUILT DATASET

A comparison with different networks proposed for fatigue detection is shown in **Table 15**. In this study, we used the LSTM network to accurately detect ATC fatigue and obtained a higher accuracy than those for the other models.

TABLE 15. Comparison of results obtained using self-built dataset.

Network	Classification scheme [1]	Feature [2]	Accuracy	Reference	Notes
Proposed network	N	F+V	0.95	—	—
	N	F	0.93	—	—
	N	V	0.92	—	—
FV-stacking	2	F+V	0.89	[38]	• Using five facial features and one vocal feature
ES-DFNN	2	E	0.89	[26]	• Only eye data was used for validation. • Use P80 as a fatigue determination criterion.
AL-based DCAE	2	V	0.92	[36]	• Only speech data was used for validation. • The feature set is based on 59 energy-based low-level descriptors.
Controller's fatigue state detection network	2	V	0.93	[69]	• Only speech data was used for validation. • Using radiotelephony communications voice.
3D-DCNN	2	F	0.94	[70]	• Only facial data was used for validation. • The network uses successive frames of pictures as input.
Non-contact drowsiness network	3 [3]	F	0.90	[71]	• Only facial data was used for validation. • The network provides a three-classification non-contact fatigue detection scheme by using PERCLOS and facial physiological signal.
Vision Transformers + YoloV5	N	F	0.93	[72]	• Only facial data was used for validation.
FaceNet + KNN	N	F	0.90	[73]	• Only facial data was used for validation.
FaceNet + SVM	N	F	0.86		
CNN	N	F	0.80	[74]	• Only facial data was used for validation. • Use the cascade object detector (Viola–Jones algorithm).
CNN	N	F	0.82	[75]	• Only facial data was used for validation. • Use the Haar cascade method.

[1] In the classification schema column, "2" represents the categorization of the fatigue levels into two classes, "3" represents the categorization of the fatigue into three classes, and "N" represents the categorization of the fatigue into more than three classes.

[2] In the feature column, "F" indicates that the network uses facial features (features containing eyes and mouth), "E" indicates that only eye features are used, and "V" indicates the use of voice features.

[3] The proposed method demonstrates dichotomous and multichotomous classifications. To adapt the method, we reclassified the SSS values and defined 1–2 as wakefulness, 3–5 as low-vigilance, and 6–7 as drowsiness.

As shown in **Table 15**, we conducted a comparative analysis of the available fatigue detection methods using a custom dataset. This comparison focused on several key aspects, including accuracy, the number of classification categories, and the specific features selected for each method.

By applying the FV-stacking network proposed by Hu et al. [38] on a self-constructed dataset, an accuracy of only 89% was obtained. Relative to the proposed recognition algorithm, the accuracy was 7% lower than that of our proposed method. This discrepancy may be due to the low dimension of the selected features.

In [26], Liang et al. proposed the ES-DFNN network to detect ATC fatigue. Using this network for the self-built dataset and employing P80 as a determinant of fatigue yielded only 89% accuracy. In contrast to our proposed network, this network only considers eye features and neglects mouth and speech features, resulting in lower accuracy.

In [36] and [69], vocal features were used as inputs to the network, achieving high accuracy by responding to the level of fatigue through speech features. Unlike the proposed networks, these networks only consider speech features and overlook facial features. Thus, the association between vocal and facial features is not considered in speech recognition, which leads to a lower accuracy. In addition, the dimension of speech features is smaller in the proposed networks, enabling faster feature extraction and recognition.

In [70] and [71], the inclusion of mouth features, in addition to eye features, marked a departure from the approach in [26], where both eye and mouth features served as inputs for the network. The use of sequential images in [70], enabled the network to assimilate contextual information, thereby enhancing accuracy. However, it's noteworthy that these networks focused exclusively on facial features, omitting vocal features from their analysis.

In [72], the incorporation of facial features into the Vision Transformers and YoloV5 network resulted in an accuracy of 93%. This outcome might stem from the fact that Vision Transformers are designed to process image inputs exclusively. Our proposed model, on the other hand, assesses fatigue levels by analyzing both image and speech features, offering a more comprehensive evaluation. Furthermore, the lower accuracy observed could also be due to a limited

TABLE 16. Comparison of results obtained using UTA-RLDD.

Network	Accuracy	Wakefulness accuracy	Low-vigilance accuracy	Drowsiness accuracy	Reference	Notes
Proposed network	0.96	0.98	—	0.95	—	No voice features
HM-LSTM network	0.65	0.81	0.32	0.82	[57]	—
LSTM network	0.61	—	—	—		
Full connection	0.57	—	—	—		
Human judgment	0.58	0.63	0.45	0.65		
Region-CNN	0.55	0.60	—	0.45	[76]	Deep learning combined with a fuzzy logic network is prominent because it prevents false alarms and achieves a specificity of 93%
Deep learning combined with fuzzy logic	0.63	0.93	—	0.35		
Vision Transformers + YoloV5	0.97	0.98	—	0.97	[72]	—
CNN + LSTM (frame level)	0.43	—	—	—	[77]	CNN + LSTM (minute segment) using KSS as label
CNN + LSTM (minute level)	0.55	—	—	—		
FaceNet + KNN	0.95	—	—	—	[73]	—
FaceNet + SVM	0.90	—	—	—		
CNN	0.96	—	—	—	[74]	CNN with transfer learning and training, and the use of the cascade object detector (Viola–Jones algorithm)
CNN	0.97	—	—	—	[75]	CNN with the Haar cascade method

sample size, which may not provide enough data for adequate regularization of the bias term in the Transformer model. Unlike this, our model does not encounter such limitations, demonstrating robust adaptability.

In [73], FaceNet was mainly used for recognizing facial feature points, with the main classification tasks being performed by KNN and SVM. KNN and SVM, unlike LSTM networks, are unable to process sequential or forward-looking information inputs, which may contribute to their relatively lower accuracy in certain applications.

In [74] and [75], the primary technique employed was CNN, with the accuracy of these methods being over 10% lower than that achieved by our proposed fatigue identification method. This lower accuracy might be attributed to the fact that the inputs were limited to facial image features, without a detailed analysis of the underlying fatigue-related digital indicators. In contrast, our proposed network delves into the fatigue-related information contained within the images and leverages the capabilities of LSTM for classification and identification, leading to improved accuracy.

Our study therefore underscores the significance of both facial and vocal features in fatigue detection. Our proposed network demonstrates superior accuracy in identifying fatigue from the same dataset compared to other models. Furthermore, we achieved a remarkable 95% accuracy in multi-category fatigue detection tasks, surpassing the performance of other networks in binary and ternary classification scenarios.

5) COMPARISON WITH DIFFERENT MULTI-CLASSIFICATION NETWORKS ON UTA-RLDD

There has been limited verification work conducted on the UTA-RLDD dataset, leading to a lack of standardized criteria, which poses some challenges. Unlike this study, most similar research projects extract different key indicators. Furthermore, while this study considers both facial and vocal features, the dataset lacks audio data. Therefore, we modified our model to exclude vocal features, focusing solely on facial features as inputs. As a result, direct comparisons with some studies may not be entirely feasible, yet such comparisons still hold interest and value. The work related to the UTA-RLDD and the findings from these analyses are summarized in **Table 16**

In [57], Ghoddoosian et al. proposed four methods to classify videos into three categories, with accuracies ranging from 58% to 65%, which are relatively low. However, the utilization of the HM-LSTM network, particularly because the inclusion of a low-vigilance category tends to diminish the overall accuracy, shows a different outcome. According to **Table 16**, the detection probabilities for both the awake and drowsy states exceed 80% when employing the HM-LSTM network. Furthermore, the introduction of additional features into the foundational LSTM network resulted in an approximate 4% improvement in the recognition rate.

Magán et al. [76] developed two models aimed at fatigue detection: one based on region-CNN and the other a combination of deep learning techniques with fuzzy logic. These models achieved accuracies of 55% and 63%, respectively.

While these accuracy levels may not seem particularly high, it is noteworthy that the prediction accuracy for identifying the awake state reached 93% with the use of the second model. The authors also highlighted the potential of deep learning methodologies for fatigue detection as a promising avenue for future research.

In [72], Krishna et al. proposed a framework using vision transformers and YoloV5 for fatigue detection, achieving a combined accuracy of 97% on UTA-RLDD. This framework also demonstrated high recognition accuracies of 98% and 97% for the awake and drowsy states, respectively. In comparison to our method, their approach achieved a 1% higher overall accuracy and a similar accuracy rate for detecting the awake state, but a 2% higher accuracy in identifying the drowsy state. These results are closely aligned with the performance of our proposed method, showcasing its strong capacity for generalization. In their model, YoloV5 is primarily employed for precise face localization, capturing facial images to feed into the Vision Transformer for subsequent learning processes. This use of YoloV5 mirrors the extraction of facial and mouth features in our method, while the Vision Transformer plays a role akin to that of the LSTM in our approach. Training the Vision Transformer with an ample dataset can mitigate the bias limitations intrinsic to the Transformer model. This process enables the model to efficiently learn and adapt to the underlying rules of bias, resulting in improved accuracy during testing phases [78]. Specifically, the extensive dataset provided by the UTA-RLDD allows the Vision Transformer to accurately discern and internalize the rules associated with the bias term, consequently achieving a remarkable accuracy rate of 97%.

In [77], Liu et al. integrated a CNN with an LSTM network for fatigue identification, utilizing the CNN for image parsing and the LSTM for time-series analysis. They developed two networks: a frame-level and a minute-level combined network, achieving accuracy rates of 43.05% and 54.71% on the UTA-RLDD, respectively. The authors noted that certain dataset features posed challenges for training, impacting model accuracy. They employed the KSS value for minute-level identification, analogous to our use of the SSS value. Contrary to their methodology, our approach did not utilize a CNN for facial feature extraction but instead used an MTCNN to analyze facial key points for fatigue feature detection. This distinction could contribute to the higher accuracy observed in our proposed model.

Adhinata et al. [73] used a CNN to extract facial features from images, employing either a multi-class SVM or a K-nearest neighbor classifier for classification. Using the multi-class SVM algorithm, they obtained an accuracy of 90%, while the K-nearest neighbor classifier reached an even higher accuracy of 95%.

In [77] and [78], the authors demonstrated that a processed CNN could adequately classify the awake and drowsy states using the UTA-RLDD, obtaining minimum accuracies of 96% and 97%, respectively.

According to **Table 16**, the accuracy of our proposed network is approximately 96%, the highest among the models listed. This accuracy is close to those of the models proposed in [73], [76], and [77]. Moreover, using the UTA-RLDD to identify and classify awake and drowsy states yielded satisfactory results, with recognition accuracies of 96% in the awake state and 95% in the drowsy state. Due to the limitations of the dataset, we could not test the performance of the vocal features part of our proposed model on this dataset; however, the results confirmed that our method exhibits high generalization capacity and accuracy based on facial features.

6) VALIDATED IN ACTUAL WORKING ENVIRONMENT

The self-built dataset was collected based on the experimental environment, which might differ from the actual working environment of controllers. To further verify the validity, practicality, and acceptability of the proposed model in the actual working environment and to confirm its generalization capacity, we collected audio and video recordings of 20 people working in their actual work environment 5 min after completing the SSS, resulting in 40 records.

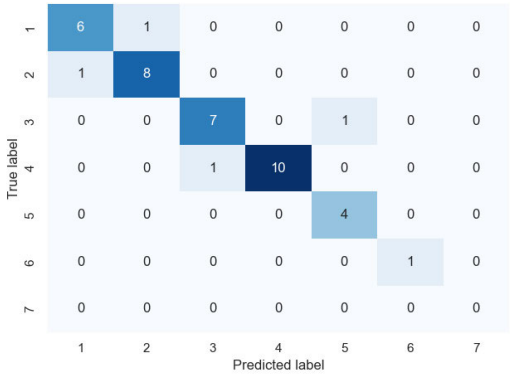


FIGURE 19. Confusion matrix for fatigue recognition in actual work environments.

A confusion matrix of 40 records of fatigue recognition in actual work environments is shown in **Fig. 19**. These 40 records were not used in the training of the proposed model. As seen in Fig. 19, the model correctly recognized fatigue in most cases, with only five records recognized incorrectly, resulting in an accuracy of 87.5%, which is considered adequate. In addition, it can be observed that the bias in recognition errors is small, with a variance of only 1.86.

V. DISCUSSION

The study demonstrated that all the features summarized in **Table 5** are reliable factors for detecting ATC fatigue. The most important aspect of this method is the ease of data acquisition without disturbing the controllers while they work, as we were able to simply obtain video and voice recordings.

We used the PERCLOS, number of yawns, average *F0*, short-time average amplitude, short-time zero-crossing rate,

HNR, jitter, shimmer, loudness, and MFCC as indicators of ATC fatigue status. Traditional fatigue prediction has often focused on one-sided feature extraction, analyzing only facial or vocal features, without effectively integrating the two. Additionally, fatigue detection has generally been limited to identifying wakefulness or fatigue without elucidating the specific degree of fatigue or classifying, recognizing, or predicting it accurately. In this study, we attempted to recognize both facial and vocal features to establish their relationships with the SSS using an LSTM network. We improved the feature dimension of fatigue by using both facial and vocal features as data for LSTM network learning and introduced the SSS to quantify fatigue levels. The results were significant, with the proposed method achieving an accuracy of 95.12%.

Using the proposed model, 21-dimensional feature vectors were extracted from the facial and vocal features and then input into the Bi-LSTM+GRU model. In the aforementioned experiments, a self-built dataset for training and testing was selected, yielding relatively accurate results. However, the dataset was relatively small. To ensure the credibility of the model, we performed five-fold cross-validation and Monte Carlo simulation testing to confirm the accuracy of the model was significant (with a 95% probability). Similar to the models described in [27], [28], [29], [57], and [70], the proposed model was developed based on the LSTM model. The difference between the proposed model and previous models lies in the increased number of extracted features. While most methods extract features of the eyes and mouth—that is, facial features—we included vocal features to obtain a more comprehensive set of features. The accuracy of most previously developed models was relatively high, especially when detecting wakefulness and drowsiness. The proposed model could also achieve an accuracy of approximately 98% when performing the above classification. However, the classification of the proposed model was not confined to just two classes.

We introduced the SSS to annotate videos. The proposed model achieved a 95% accuracy rate under multi-classification conditions, indicating some improvement compared to the previously proposed version. Similar to the research presented in [34], [35], [70], and [71], we used the LSTM network as the classification model. Moreover, as shown in **Table 15**, when comparing several fatigue detection networks, our proposed model exhibited the highest accuracy and outperformed other centralized methods. This superiority may stem from the LSTM better performance of the network on data with time-series characteristics and the utilization of both facial and voice features as inputs to the network. Moreover, we tested the proposed model in an actual working environment to assess its validity, practicability, and acceptability. We verified that despite differences between the dataset and the actual working environment, the proposed model could identify the fatigue level of the controller more accurately, demonstrating the high generalization capacity of the model.

Additionally, in this study, similar to the studies reported in [23], [37], and [70], the fatigue scale was introduced as a label value for model learning. Previous authors divided the test values of the scale into intervals to form fatigue and wakefulness intervals before inputting this information into their proposed models. In contrast, in this study, we manually divided the scale values of fatigue and wakefulness first, then directly performed model training and testing without dividing the scale values. Both of these approaches achieved good results, which holds specific significance for further determining the level of ATC fatigue. In summary, the proposed model combined facial and vocal features, used the SSS to evaluate the ATC fatigue level, and inputted the combined features into the LSTM model to predict the refined ATC fatigue level.

However, this experiment had several limitations.

(1) The sample size used to obtain the data was too small, which could have led to poor generalization of the model and less comprehensive learning of the LSTM network, resulting in low prediction accuracy.

(2) The clarity of the video capture, the angle and distance of the camera, and equipment errors were not considered. Consequently, the collected data may have been inaccurate, leading the model to be imperfect.

(3) Differences existed among the individual subjects, such as gender and phonetic differences, which were not considered in this experiment.

(4) The sample data collection was conducted in an experimental environment, which may differ from the actual work environment of the controller, potentially leading to inaccurate identification of fatigue levels.

VI. CONCLUSION

In this study, we proposed a fatigue detection method that leverages facial and vocal features. We chose an LSTM-based model because both face and voice tones convey significant information about fatigue levels. The aim of our study was to explore the link between ATC fatigue and these facial-vocal features. We synchronously collected self-fatigue assessment information using the SSS, PERCLOS, number of yawns, average *F0*, short-time average amplitude, short-time zero-crossing rate, HNR, jitter, shimmer, loudness, and MFCC. Training the LSTM (Bi-LSTM+GRU) network enabled us to predict ATC fatigue states, demonstrating that combining facial and vocal features enhances fatigue identification. We obtained a maximum accuracy rate of 95.12%, providing a theoretical basis for the use of combined facial and vocal features in detecting fatigue.

The proposed method can be applied for real-time applications in preventing air traffic incidents caused by ATC fatigue. This method stands out for its straightforward data collection, minimal interference with ATC operations, and cost-effectiveness. Furthermore, our research showcases the efficiency of a non-invasive approach to fatigue detection, which has substantial practical implications.

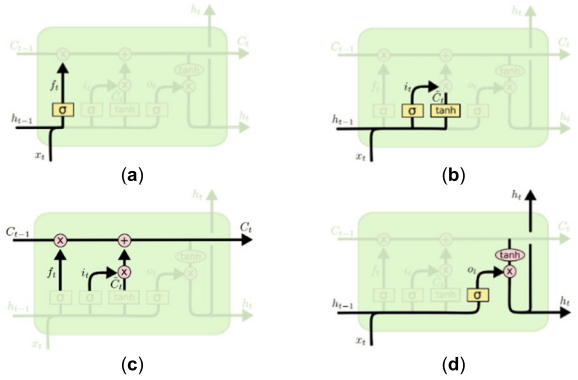


FIGURE 20. LSTM steps: (a) Oblivion, (b) previous input, (c) update cell status, and (d) output.

APPENDIX
LSTM INTRODUCTION

The structure of LSTM cells is more complex than that of RNN cells. Each LSTM cell comprises three *sigmoid* layers and one *tanh* layer, facilitating the updating or forgetting of relevant information—a primary innovation of the LSTM network. The traditional LSTM network can be divided into four steps, as shown in **Fig. 20.**

During information processing, LSTM cells discard some information in the first step to simulate the human tendency to forget and disregard certain information during cognitive processes (**Fig. 20(a)**). In this process, the output (h_{t-1}) from the previous step and the input (x_t) at the current time are nonlinearly mapped through the *sigmoid* layer, resulting in the output f_t, as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f). \tag{A1}$$

Simultaneously, the input undergoes processing through the input gate (**Fig. 20(b)**):

- A part of the input value is updated through the *sigmoid* layer to obtain the output i_t;
- The remaining part is inputted into the tanh layer to create a new candidate vector ($\tilde{C}_t$) for subsequent calculations.

The above steps can be expressed as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{A2}$$

and

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \tag{A3}$$

By incorporating *forgetting* and *input* mechanisms, the updated data (as shown in **Fig. 20(c)**) can be expressed as follows:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t. \tag{A4}$$

Finally, the filtered output is determined by the process shown in **Fig. 20(d)**, which can be expressed as follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{A5}$$

$$h_t = o_t * \tanh(C_t). \tag{A6}$$

Based on this concept and leveraging the advantages of LSTM networks, we propose an LSTM-based fatigue detection method in this paper. The flowchart of the ATC fatigue detection method is shown in **Fig. 20**.

REFERENCES

[1] *Working Document for the Aviation System Block Upgrades: The Framework for Global Harmonization*, Int. Civil Aviation Org., Montreal, QC, Canada, 2016.
[2] *Air Traffic Management*, Int. Civil Aviation Org., Montreal, QC, Canada, 2016.
[3] S. Bendak and H. S. J. Rashid, "Fatigue in aviation: A systematic review of the literature," *Int. J. Ind. Ergonom.*, vol. 76, Mar. 2020, Art. no. 102928, doi: 10.1016/j.ergon.2020.102928.
[4] T. Akerstedt, "Consensus statement: Fatigue and accidents in transport operations," *J. Sleep Res.*, vol. 9, no. 4, p. 395, Dec. 2000.
[5] *Manual for the Oversight of Fatigue Management Approaches*, Int. Civil Aviation Org., Montreal, QC, Canada, 2020.
[6] *Fatigue Management Guide for Airline Operators*, Int. Civil Aviation Org., Montreal, QC, Canada, 2015.
[7] *Fatigue Management Guide for General Aviation Operators of Large and Turbojet Aeroplanes*, Int. Civil Aviation Org., Montreal, QC, Canada, 2016.
[8] *Fatigue Management Guide for Air Traffic Service Providers*, Int. Civil Aviation Org., Montreal, QC, Canada, 2016.
[9] *Fatigue Management Guide for Helicopter Operators*, Int. Civil Aviation Org., Montreal, QC, Canada, 2020.
[10] D. F. Dinges, "An overview of sleepiness and accidents," *J. Sleep Res.*, vol. 4, no. 2, pp. 4–14, Dec. 1995, doi: 10.1111/j.1365-2869.1995.tb00220.x.
[11] L. S. Aaronson, C. S. Teel, V. Cassmeyer, G. B. Neuberger, L. Pallikkathayil, J. Pierce, A. N. Press, P. D. Williams, and A. Wingate, "Defining and measuring fatigue," *Image, J. Nursing Scholarship*, vol. 31, no. 1, pp. 45–50, Mar. 1999, doi: 10.1111/j.1547-5069.1999.tb00420.x.
[12] M.-L. Chen, S.-Y. Lu, and I.-F. Mao, "Subjective symptoms and physiological measures of fatigue in air traffic controllers," *Int. J. Ind. Ergonom.*, vol. 70, pp. 1–8, Mar. 2019, doi: 10.1016/j.ergon.2018.12.004.
[13] S. Ahn, T. Nguyen, H. Jang, J. G. Kim, and S. C. Jun, "Exploring neuro-physiological correlates of drivers' mental fatigue caused by sleep deprivation using simultaneous EEG, ECG, and fNIRS data," *Frontiers Hum. Neurosci.*, vol. 10, p. 219, May 2016, doi: 10.3389/fnhum.2016.00219.
[14] J. A. Horne and O. Ostberg, "A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms," *Int. J. Chronobiol.*, vol. 4, pp. 97–110, Jan. 1976.
[15] T. Åkerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *Int. J. Neurosci.*, vol. 52, nos. 1–2, pp. 29–37, Jan. 1990, doi: 10.3109/00207459008994241.
[16] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, "Stanford sleepiness scale (SSS)," in *STOP, THAT and One Hundred Other Sleep Scales*, A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, Eds. New York, NY, USA: Springer, 2011, pp. 369–370, doi: 10.1007/978-1-4419-9893-4_91.
[17] T. Chalder, G. Berelowitz, T. Pawlikowska, L. Watts, S. Wessely, D. Wright, and E. P. Wallace, "Development of a fatigue scale," *J. Psychosomatic Res.*, vol. 37, no. 2, pp. 147–153, Feb. 1993, doi: 10.1016/0022-3999(93)90081-p.
[18] L. Colligan, H. W. W. Potts, C. T. Finn, and R. A. Sinkin, "Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record," *Int. J. Med. Informat.*, vol. 84, no. 7, pp. 469–476, Jul. 2015, doi: 10.1016/j.ijmedinf.2015.03.003.
[19] W. H. Gu, Y. Zhu, X. D. Chen, L. F. He, and B. B. Zheng, "Hierarchical CNN-based real-time fatigue detection system by visual-based technologies using MSP model," *IET Image Process.*, vol. 12, no. 12, pp. 2319–2329, Dec. 2018, doi: 10.1049/iet-ipr.2018.5245.
[20] T. T. Truong, D. Dinh-Cong, J. Lee, and T. Nguyen-Thoi, "An effective deep feedforward neural networks (DFNN) method for damage identification of truss structures using noisy incomplete modal data," *J. Building Eng.*, vol. 30, Jul. 2020, Art. no. 101244, doi: 10.1016/j.jobe.2020.101244.
[21] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306, doi: 10.1016/j.physd.2019.132306.

[22] W. W. Wierwille, S. S. Wreggit, C. L. Kirn, L. A. Ellsworth, and R. J. Fairbanks, "Research on vehicle-based driver status/performance monitoring; Development, validation, and refinement of algorithms for detection of driver drowsiness," U.S. Dept. Transp., Washington, DC, USA, Tech. Rep. DOT HS 808 247, 1994.
[23] D. Sommer and M. Golz, "Evaluation of PERCLOS based current fatigue monitoring technologies," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol.*, Aug. 2010, pp. 4456–4459, doi: 10.1109/IEMBS.2010.5625960.
[24] Z. Zhao, N. Zhou, L. Zhang, H. Yan, Y. Xu, and Z. Zhang, "Driver fatigue detection based on convolutional neural networks using EM-CNN," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–11, Nov. 2020, doi: 10.1155/2020/7251280.
[25] R. Li, Y. V. Chen, and L. Zhang, "A method for fatigue detection based on driver's steering wheel grip," *Int. J. Ind. Ergonom.*, vol. 82, Mar. 2021, Art. no. 103083, doi: 10.1016/j.ergon.2021.103083.
[26] H. Liang, C. Liu, K. Chen, J. Kong, Q. Han, and T. Zhao, "Controller fatigue state detection based on ES-DFNN," *Aerospace*, vol. 8, no. 12, p. 383, Dec. 2021, doi: 10.3390/aerospace8120383.
[27] Y. Zhao, K. Xie, Z. Zou, and J.-B. He, "Intelligent recognition of fatigue and sleepiness based on InceptionV3-LSTM via multi-feature fusion," *IEEE Access*, vol. 8, pp. 144205–144217, 2020, doi: 10.1109/ACCESS.2020.3014508.
[28] L. Chen, G. Xin, Y. Liu, and J. Huang, "Driver fatigue detection based on facial key points and LSTM," *Secur. Commun. Netw.*, vol. 2021, pp. 1–9, Jun. 2021, doi: 10.1155/2021/5383573.
[29] B. Akrout and S. Fakhfakh, "How to prevent drivers before their sleepiness using deep learning-based approach," *Electronics*, vol. 12, no. 4, p. 965, Feb. 2023, doi: 10.3390/electronics12040965.
[30] J. Wang, Y. Xu, J. Tian, H. Li, W. Jiao, Y. Sun, and G. Li, "Driving fatigue detection with three non-hair-bearing EEG channels and modified transformer model," *Entropy*, vol. 24, no. 12, p. 1715, Nov. 2022, doi: 10.3390/e24121715.
[31] S. Kumar, M. K. Chaube, S. H. Alsamhi, S. K. Gupta, M. Guizani, R. Gravina, and G. Fortino, "A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques," *Comput. Methods Programs Biomed.*, vol. 226, Nov. 2022, Art. no. 107109, doi: 10.1016/j.cmpb.2022.107109.
[32] X. Yu, C.-H. Chen, and H. Yang, "Air traffic controllers' mental fatigue recognition: A multi-sensor information fusion-based deep learning approach," *Adv. Eng. Informat.*, vol. 57, Aug. 2023, Art. no. 102123, doi: 10.1016/j.aei.2023.102123.
[33] S. Milosevic, "Drivers' fatigue studies," *Ergonomics*, vol. 40, no. 3, pp. 381–389, Mar. 1997, doi: 10.1080/001401397188215.
[34] X. Li, N. Tan, T. Wang, and S. Su, "Detecting driver fatigue based on nonlinear speech processing and fuzzy SVM," in *Proc. 12th Int. Conf. Signal Process. (ICSP)*, Zhejiang, China, Oct. 2014, pp. 510–515, doi: 10.1109/ICOSP.2014.7015057.
[35] C. Craye, A. Rashwan, M. S. Kamel, and F. Karray, "A multi-modal driver fatigue and distraction assessment system," *Int. J. Intell. Transp. Syst. Res.*, vol. 14, no. 3, pp. 173–194, Sep. 2016, doi: 10.1007/s13177-015-0112-9.
[36] Z. Shen and Y. Wei, "A high-precision feature extraction network of fatigue speech from air traffic controller radiotelephony based on improved deep learning," *ICT Exp.*, vol. 7, no. 4, pp. 403–413, Dec. 2021, doi: 10.1016/j.icte.2021.01.002.
[37] X. Gao, K. Ma, H. Yang, K. Wang, B. Fu, Y. Zhu, X. She, and B. Cui, "A rapid, non-invasive method for fatigue detection based on voice information," *Frontiers Cell Develop. Biol.*, vol. 10, Sep. 2022, Art. no. 994001, doi: 10.3389/fcell.2022.994001.
[38] Y. Hu, Z. Liu, A. Hou, C. Wu, W. Wei, Y. Wang, and M. Liu, "On fatigue detection for air traffic controllers based on fuzzy fusion of multiple features," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–10, Oct. 2022, doi: 10.1155/2022/4911005.
[39] Q. Zhan, W. Zhou, J. Gao, W. Li, and X. Zhang, "A review of driver fatigue detection and warning based on multi-information fusion," in *Proc. 3rd Int. Forum Connected Automated Vehicle Highway Syst. Through China Highway Transp. Soc.*, Dec. 2020, p. 5143, doi: 10.4271/2020-01-5143.
[40] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
[41] H. He, H. Xu, Y. Zhang, K. Gao, H. Li, L. Ma, and J. Li, "Mask R-CNN based automated identification and extraction of oil well sites," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102875, doi: 10.1016/j.jag.2022.102875.
[42] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.
[43] D. Zhang, J. Li, and Z. Shan, "Implementation of dlib deep learning face recognition technology," in *Proc. Int. Conf. Robots Intell. Syst. (ICRIS)*, Sanya, China, Nov. 2020, pp. 88–91, doi: 10.1109/ICRIS52159.2020.00030.
[44] S. Sharma, K. Shanmugasundaram, and S. K. Ramasamy, "FAREC—CNN based efficient face recognition technique using dlib," in *Proc. Int. Conf. Adv. Commun. Control Comput. Technol. (ICACCCT)*, May 2016, pp. 192–195, doi: 10.1109/ICACCCT.2016.7831628.
[45] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Guang Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A framework for building perception pipelines," 2019, *arXiv:1906.08172*.
[46] W. Yaqub, M. Mohanty, and B. Suleiman, "Image-hashing-based anomaly detection for privacy-preserving online proctoring," 2021, *arXiv:2107.09373*.
[47] J. Cech and T. Soukupova, "Real-time eye blink detection using facial landmarks," presented at the 21st Comput. Vis. Winter Workshop Rimske Toplice, Slovenia, Feb. 2016.
[48] J. Schoentgen, "Stochastic models of jitter," *J. Acoust. Soc. Amer.*, vol. 109, no. 4, pp. 1631–1650, Apr. 2001, doi: 10.1121/1.1350557.
[49] N. Sripriya, S. Poornima, R. Shivaranjani, and P. Thangaraju, "Non-intrusive technique for pathological voice classification using jitter and shimmer," in *Proc. Int. Conf. Comput., Commun. Signal Process. (ICCCSP)*, Jan. 2017, pp. 1–6, doi: 10.1109/ICCCSP.2017.7944104.
[50] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," 2010, *arXiv:1003.4083*.
[51] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Commun.*, vol. 54, no. 4, pp. 543–565, May 2012, doi: 10.1016/j.specom.2011.11.004.
[52] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. SPECOM*, 2005, pp. 191–194.
[53] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Inf. Media Technol.*, vol. 2, no. 3, 835–848, 2007, doi: 10.11185/imt.2.835.
[54] O. K. Hamid, "Frame blocking and windowing speech signal," *J. Inf., Commun., Intell. Syst. (JICIS)*, vol. 4, no. 5, pp. 87–94, 2018.
[55] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[56] A. C. Gallup, A. M. Church, and A. J. Pelegrino, "Yawn duration predicts brain weight and cortical neuron number in mammals," *Biol. Lett.*, vol. 12, no. 10, Oct. 2016, Art. no. 20160545, doi: 10.1098/rsbl.2016.0545.
[57] R. Ghoddoosian, M. Galib, and V. Athitsos, "A realistic dataset and baseline temporal model for early drowsiness detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 178–187. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/AMFG/Ghoddoosian_A_Realistic_Dataset_and_Baseline_Temporal_Model_for_Early_Drowsiness_CVPRW_2019_paper.html
[58] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[59] S. Li, W. Luan, C. Wang, Y. Chen, and Z. Zhuang, "Degradation prediction of proton exchange membrane fuel cell based on Bi-LSTM-GRU and ESN fusion prognostic framework," *Int. J. Hydrogen Energy*, vol. 47, no. 78, pp. 33466–33478, Sep. 2022, doi: 10.1016/j.ijhydene.2022.07.230.
[60] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110212, doi: 10.1016/j.chaos.2020.110212.
[61] M. Konstantinou, S. Peratikou, and A. G. Charalambides, "Solar photovoltaic forecasting of power output using LSTM networks," *Atmosphere*, vol. 12, no. 1, p. 124, Jan. 2021, doi: 10.3390/atmos12010124.
[62] T.-T. Wong and P.-Y. Yeh, "Reliable accuracy estimates from K-fold cross validation," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 8, pp. 1586–1594, Aug. 2020, doi: 10.1109/TKDE.2019.2912815.
[63] P. Liashchynskyi and P. Liashchynskyi, "Grid search, random search, genetic algorithm: A big comparison for NAS," 2019, *arXiv:1912.06059*.

[64] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika*, vol. 14, no. 4, p. 1502, Dec. 2016.
[65] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, "Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm," *J. Rock Mech. Geotechnical Eng.*, vol. 14, no. 4, pp. 1292–1303, Aug. 2022, doi: 10.1016/j.jrmge.2022.03.002.
[66] M. Beer and P. D. Spanos, "Neural network based Monte Carlo simulation of random processes," in *Proc. Int. Conf. Struct. Saf. Rel.*, 2005, pp. 2179–2186. [Online]. Available: http://rcswww.urz.tu-dresden.de/~statik/sfb/download/e4_icossar2005_pdf_017.pdf
[67] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*. Hoboken, NJ, USA: Wiley, 2016.
[68] Y. Zhang, R. Xiong, H. He, and M. G. Pecht, "Long short-term memory recurrent neural network for remaining useful life prediction of lithium-ion batteries," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 5695–5705, Jul. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8289406
[69] N. Wu and J. Sun, "Fatigue detection of air traffic controllers based on radiotelephony communications and self-adaption quantum genetic algorithm optimization ensemble learning," *Appl. Sci.*, vol. 12, no. 20, p. 10252, Oct. 2022, doi: 10.3390/app122010252.
[70] J. Yu, S. Park, S. Lee, and M. Jeon, "Driver drowsiness detection using condition-adaptive representation learning framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 11, pp. 4206–4218, Nov. 2019, doi: 10.1109/TITS.2018.2883823.
[71] R. C.-H. Chang, C.-Y. Wang, W.-T. Chen, and C.-D. Chiu, "Drowsiness detection system based on PERCLOS and facial physiological signal," *Sensors*, vol. 22, no. 14, p. 5380, Jul. 2022, doi: 10.3390/s22145380.
[72] G. Sai Krishna, K. Supriya, J. Vardhan, and M. Rao, "Vision transformers and YoloV5 based driver drowsiness detection framework," 2022, *arXiv:2209.01401*.
[73] F. D. Adhinata, D. P. Rakhmadani, and D. Wijayanto, "Fatigue detection on face image using FaceNet algorithm and K-nearest neighbor classifier," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 7, no. 1, pp. 22–30, 2021.
[74] I. Nasri, M. Karrouchi, H. Snoussi, K. Kassmi, and A. Messaoudi, "Detection and prediction of driver drowsiness for the prevention of road accidents using deep neural networks techniques," in *WITS 2020* (Lecture Notes in Electrical Engineering), S. Bennani, Y. Lakhrissi, G. Khaissidi, A. Mansouri, and Y. Khamlichi, Eds. Singapore: Springer, 2022, pp. 57–64, doi: 10.1007/978-981-33-6893-4_6.
[75] R. Tamanani, R. Muresan, and A. Al-Dweik, "Estimation of driver vigilance status using real-time facial expression and deep learning," *IEEE Sensors Lett.*, vol. 5, no. 5, pp. 1–4, May 2021, doi: 10.1109/LSENS.2021.3070419.
[76] E. Magán, M. P. Sesmero, J. M. Alonso-Weber, and A. Sanchis, "Driver drowsiness detection by applying deep learning techniques to sequences of images," *Appl. Sci.*, vol. 12, no. 3, p. 1145, Jan. 2022, doi: 10.3390/app12031145.
[77] P. Liu, H.-L. Chi, X. Li, and J. Guo, "Effects of dataset characteristics on the performance of fatigue detection for crane operators using hybrid deep neural networks," *Autom. Construction*, vol. 132, Dec. 2021, Art. no. 103901, doi: 10.1016/j.autcon.2021.103901.
[78] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

ZHOUSHENG HUANG received the B.Eng. degree in traffic and transportation from the Civil Aviation Flight University of China, Sichuan, China, in 2020, where he is currently pursuing the M.Eng. degree. Since 2020, he has been employed as a Teaching Assistant with the Civil Aviation Flight University of China. His main research interests include air traffic management and artificial intelligence.

WEIZHEN TANG received the B.Eng. degree in traffic and transportation from the Civil Aviation Flight University of China, Sichuan, China, in 2000. In the same year, he joined the Civil Aviation Flight University of China, where he is currently a Professor. His main research interests include air traffic management and safety management.

QIQI TIAN is currently pursuing the M.Eng. degree in transportation with the Civil Aviation Flight University of China, Sichuan, China. Her main research interest includes air traffic management.

TING HUANG received the B.Eng. degree in marine engineering from Chongqing Jiaotong University, in 2022. She is currently pursuing the M.Eng. degree with the Civil Aviation Flight University of China. Her main research interest includes air traffic management.

JINZE LI received the B.Eng. degree in traffic and transportation from the Nanhang Jincheng College, Nanjing, China, in 2018. He is currently pursuing the master's degree with the Civil Aviation Flight University of China. His main research interest includes air traffic management.