Hindawi
Computational and Mathematical Methods in Medicine
Volume 2022, Article ID 4911005, 10 pages
https://doi.org/10.1155/2022/4911005

Research Article

On Fatigue Detection for Air Traffic Controllers Based on Fuzzy Fusion of Multiple Features

Yi Hu,[1] **Zhuo Liu**,[2] **Aiqin Hou**,[2] **Chase Wu**,[3] **Wenbin Wei**,[4] **Yanjun Wang**,[1] **and Min Liu**[5]

[1]*College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu 210016, China*
[2]*School of Information Science and Technology, Northwest University, Xi'an, Shaanxi 710127, China*
[3]*Department of Data Science, New Jersey Institute of Technology, Newark, NJ 07102, USA*
[4]*Department of Aviation and Technology, San Jose State University, San Jose, CA 95192, USA*
[5]*Zhongke Haoyin Intelligent Technology Co., Ltd., Hefei, Anhui 230088, China*

Correspondence should be addressed to Aiqin Hou; houaiqin@nwu.edu.cn

Received 19 May 2022; Revised 1 September 2022; Accepted 12 September 2022; Published 11 October 2022

Academic Editor: A. S. Albahri

Fatigue detection for air traffic controllers is an important yet challenging problem in aviation safety research. Most of the existing methods for this problem are based on facial features. In this paper, we propose an ensemble learning model that combines both facial features and voice features and design a fatigue detection method through multifeature fusion, referred to as Facial and Voice Stacking (FV-Stacking). Specifically, for facial features, we first use OpenCV and Dlib libraries to extract mouth and eye areas and then employ a combination of M-Convolutional Neural Network (M-CNN) and E-Convolutional Neural Network (E-CNN) to determine the state of mouth and eye closure based on five features, i.e., blinking times, average blinking time, average blinking interval, Percentage of Eyelid Closure over the Pupil over Time (PERCLOS), and Frequency of Open Mouth (FOM). For voice features, we extract the Mel-Frequency Cepstral Coefficients (MFCC) features of speech. Such facial features and voice features are fused through a carefully designed stacking model for fatigue detection. Real-life experiments are conducted on 14 air traffic controllers in Southwest Air Traffic Management Bureau of Civil Aviation of China. The results show that the proposed FV-Stacking method achieves a detection accuracy of 97%, while the best accuracy achieved by a single model is 92% and the best accuracy achieved by the state-of-the-art detection methods is 88%.

1. Introduction

The 2006-2015 statistics of flight incidents in China broken down by causes show that human factors account for 25.67% [1]. From 1994 to 2020, there are 97 accidents in China caused by air traffic controllers [2]. Also, a survey on controller fatigue from National Transportation Safety Board (NTSB) shows that a number of major investigations identify fatigue as a probable cause, contributing factor, or a finding [3]. The above studies and statistics have made it very clear that timely and accurate fatigue detection for air traffic controllers performing control and command operations on site is critical to minimizing aviation safety hazards. Fatigue detection is mainly divided into subjective detection and objective detection. Subjective detection is to classify and quantify the fatigue status based on the subject's subjective performance. The detection methods in this category have the advantages of convenient operation and low cost, and have been widely adopted. However, they also suffer from some disadvantages such as poor real-time performance and the impact of human subjective consciousness. Subjective detection methods can be further divided into three subcategories: questionnaires and subjective evaluation forms, oral question-and-answer analysis, and active detection models. Lee and Kim developed hypotheses and survey questions based on interviews with 929 pilots and conducted a nationwide survey. They concluded that inadequate planning operation, flight direction, culturally different partnership, aircraft environment, job assignment,

racial difference, hotel environment, and other factors can cause pilot fatigue [4]. Jiang et al. designed a questionnaire based on Theory of Planned Behavior (TPB), which effectively reveals the psychological factors related to fatigue driving [5].

Objective detection has the advantages of high accuracy and reliability, and it is not affected by the subject's subjective consciousness. The methods in this category have become the focus of research for fatigue detection and can be divided into the following two subcategories: contact and noncontact. Most traditional contact-based methods for fatigue detection measure physiological signals such as electrocardiograms and brain waves [6, 7]. Such methods through body contact are able to yield high detection accuracy but may interfere with the normal operation of air traffic controllers. Most noncontact-based methods mainly track facial expressions, such as mouth state detection [8], eye tracking [9, 10], and reaction time [11]. Verma et al. proposed to detect fatigue by comparing the location of the joints of the current posture [12]. Since noncontact methods are noninvasive and easy to instrument, they have received a great deal of attention.

In this paper, we propose a fatigue detection method through multifeature fusion based on ensemble learning, referred to as Facial and Voice Stacking (FV-Stacking). Specifically, for facial features, we first use OpenCV and Dlib libraries to extract mouth and eye areas and then employ a combination of M-Convolutional Neural Network (M-CNN) and E-Convolutional Neural Network (E-CNN) to determine the state of mouth and eye closure based on five features, i.e., blinking times, average blinking time, average blinking interval, Percentage of Eyelid Closure over the Pupil over Time (PERCLOS), and Frequency of Open Mouth (FOM). For voice features, we extract the Mel-Frequency Cepstral Coefficients (MFCC) features of speech. Such facial features and voice features are fused using a carefully designed stacking model for fatigue detection. The stacking framework increases the amount of information used for ensemble learning and can integrate different types of features by choosing and stacking appropriate base models. Real-life experiments are conducted on 14 air traffic controllers in Southwest Air Traffic Management Bureau of Civil Aviation of China. The experimental results show that the proposed FV-Stacking method achieves a detection accuracy of 97%, while the best accuracy achieved by a single model is 92%, and the best accuracy achieved by the state-of-the-art detection methods is 88%. The main contributions of our work are summarized as follows:

(1) We develop a machine learning model that can recognize the closed state of the mouth and eyes with high accuracy

(2) We fuse speech features and facial features to detect the fatigue state of air traffic controllers

(3) We design an FV-Stacking ensemble learning model and achieve an accuracy rate of 97% for fatigue detection

The rest of this paper is organized as follows. Section 2 conducts a survey of related work. Section 3 details the design of the proposed ensemble learning model through multifeature fusion. Section 4 presents and analyzes experimental results. We conclude our work in Section 5.

2. Related Work

Fatigue detection is used in various scenarios and is mainly divided into two categories, i.e., subjective detection and objective detection.

In subjective methods, fatigue ranges are often used. Williamson et al. [13] systematically studied the impact of lack of sleep on fatigue and established a set of subjective methods that can be used to assess fatigue. De Vries et al. [14] claimed that the Fatigue Assessment Scale is the most promising fatigue measure, where workers are requested to fill out questionnaires before and after work to divide the fatigue scale.

Among objective methods, there are contact methods and noncontact methods for fatigue detection, depending on whether or not the testing tool needs to physically touch the tested person during testing. Heart rates, brain waves, and electrocardiograms (EEG) are often used as common indicators in contact-based detection methods. Arnau et al. [15] used EEG to study the relationship between mental fatigue and age. Murugan et al. [7] extracts 13 electrocardiogram (ECG) signal features and classifies them through machine learning to determine the fatigue status of a person. Chen et al. [16] determined whether or not the air traffic controller is fatigued by measuring physiological information including flicker fusion threshold, thumb/index finger strength, and systolic and diastolic pressure before and after work. For noncontact detection, many methods consider facial expressions and voice signals. Dinges and Grace [17] proposed PERCLOS, a physical quantity measuring fatigue/drowsiness, which is defined as a certain percentage (e.g., 70% or 80%) of time when the eyes are closed per unit time. Generally, a tested person is considered to be fatigued if PERCLOS exceeds a certain threshold. Zhang et al. [18] used a convolutional neural network to determine the closure status of eyes, calculated PERCLOS based on this, and combined the number of blinks per unit time to identify the fatigue state. Nie et al. [19] used PERCLOS, blink rate, and eye closure time to detect fatigue status. Gu et al. [20] detected fatigue status by calculating PERCLOS and yawn frequency. Similarly, voice features are also considered in some noncontact methods for fatigue detection. Shen et al. [21] used Revised Fractal Dimension Feature to determine the fatigue status of air traffic controllers.

3. Ensemble Learning Thorough Multifeature Fusion

Ensemble learning accomplishes the learning task by combining multiple models. The selection of an ensemble learning model follows the principle of "good but different". It uses a series of base models and some rules to integrate multiple learning results to obtain a final one, which is expected to outperform a single learning method [22]. Ensemble learning includes several schemes, these are Bagging, Stacking, Boosting, Blending, etc.

We propose a model of FV-Stacking to combine facial features and voice features for fatigue detection. Stacking is a layered model integration framework. The first layer is composed of multiple base learners. In the FV-Stacking framework, we combine five base models, i.e., Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN), which take the original training dataset as input. The second layer is a simple LR model, which takes the output of the base learners in the first layer as input. The architecture of the proposed FV-Stacking framework is shown in Figure 1.

A brief introduction to each of the base models used in FV-Stacking is provided as follows.

(1) *LR*. Logistic regression uses the logistic function sigmoid to map the result of linear regression to the range of [0,1]. In FV-Stacking, LR is used in both the first and the second layers. In the first layer, logistic regression classifies fatigue status based on facial features, while in the second layer, it is used to classify the combined inputs of all base models

(2) *SVM*. Support vector machine is a classification model, and its basic model is a linear classifier with the largest interval defined in the feature space. The key idea is to solve the separating hyperplane that can correctly divide the training dataset and have the largest geometric interval. In this paper, we use a linear support vector machine to recognize facial features

(3) *DT*. Decision tree is a supervised machine learning algorithm based on a tree structure, in which each internal node represents a judgement of an attribute, each branch represents the result of a judgement, and each leaf node represents a classification method. In this paper, we use CART decision tree to recognize facial features

(4) *LSTM*. Long Short-Term Memory is a recurrent neural network and is well-suited to classify time series data. In this paper, LSTM is used to process MFCC features, as illustrated in Figure 2

(5) *CNN*. Convolutional neural network is a type of neural network that performs convolution calculation and has a deep structure. It includes multiple layers including convolutional layer, pooling layer, and fully connected layer. The convolutional layer and the pooling layer perform feature extraction on the input data, and the fully connected layer performs a nonlinear combination of the extracted features to obtain the output. In this paper, CNN is used to process MFCC features, using three convolutional layers, three pooling layers, a flatten operation, one fully connected layer, and one sigmoid classifier, as illustrated in Figure 3

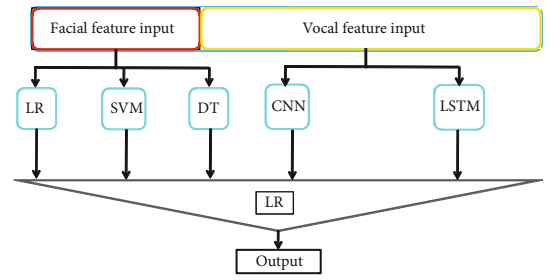


Figure 1: FV-Stacking architecture.

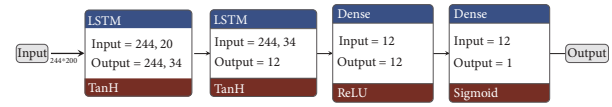


Figure 2: LSTM structure.

The input of FV-Stacking includes visual data and voice data. The facial feature input includes blinking times, average blinking time, average blinking interval, PERCLOS, and FOM. The voice feature input is the MFCC feature.

3.1. Facial Feature Extraction

3.1.1. Face Detection and Feature Point Extraction. Face detection is to determine face images, and feature point extraction is to identify feature points in face images. These are the most critical steps in facial recognition. The quality of a detected face and the accuracy of the feature point location directly affect the results of subsequent processes. In this paper, we use Dlib Library to detect faces and extract facial feature points. Dlib is a modern C++ toolkit that includes a variety of machine learning algorithms and tools, providing high-quality machine learning, image processing, deep learning, and face recognition library [23]. Face recognition algorithms include face detection, face feature extraction, and face feature vector calculation. Hence, we choose the Dlib library to implement a high-quality face recognition system. In the Dlib library, the pretrained facial landmark detector is used to estimate the location of 68 coordinates (x, y) that map to facial structures on the face. For illustration, the indexes of the 68 coordinates are visualized in Figure 4. In this paper, we employ the Dlib library to extract the 68 coordinates of the face and locate the eye and the mouth.

3.1.2. Eye Closure Status Recognition. After extracting 68 eye detection points, we use these points to construct an eye area of size 32×26 based on the facial landmarks defined in Eq. (1), as illustrated in Figure 5. Figure 5(a) is the marked points of a human eye in the video, and Figure 5(b) is the extracted grayscale image of the human eye.

$$w_e = 1.2 * X,$$
$$H_e = \frac{w_e * 34}{26}. \quad (1)$$

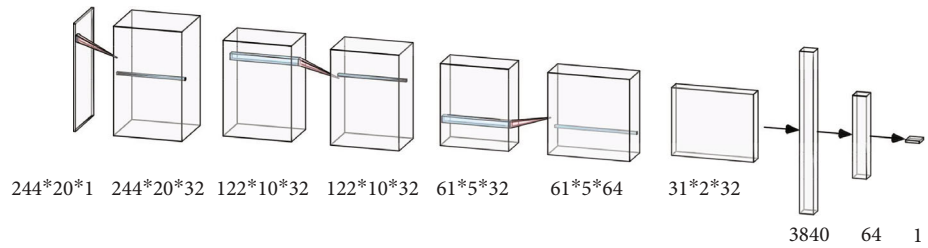


Figure 3: CNN structure.

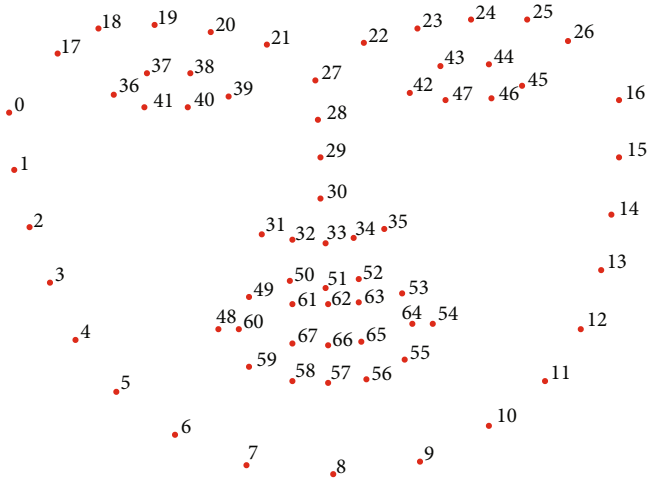


Figure 4: The location of 68 feature points on a face.

Once the eye area is identified, we use E-CNN to determine the eye's closure status [24]. E-CNN contains three convolutional layers, three pool layers, two fully connected layers, a flatten operation, and one sigmoid classifier, as illustrated in Figure 6 [25]. The input is a grayscale image of size $26 \times 34 \times 1$.

3.1.3. Mouth Closure Status Recognition. Among 68 facial detection points, we use those points of mouth to extract the mouth area of size 120×80, based on the facial landmarks defined in Eq. (2). An extraction result is plotted in Figure 7 for illustration. Figure 7(a) is the marked points of an air traffic controller's mouth in the video, and Figure 7(b) is the extracted grayscale image of the air traffic controller's mouth.

$$W_m = 1.2 * Y_m,$$
$$H_m = \frac{w_m * 120}{80}. \quad (2)$$

Once the mouth area is identified, we use M-CNN to determine the mouth's closure status [26]. Convolutional neural networks contain three convolutional layers, three pool layers, two fully connected layers, a flatten operation, and one sigmoid classifier, as illustrated in Figure 8.

The input is a grayscale image of size $80 \times 120 \times 1$.

3.1.4. Eye and Mouth Features. We generated two queues when identifying air traffic controllers in the video stream with M-CNN and E-CNN. As shown in Figure 9, the first queue stores the detection results of eye state, and the second queue stores the detection results of mouth state. We use a flag number to represent the closure state of the eye or mouth in each frame: the flag '1' indicates that the eye or mouth is open, and the flag '0' indicates that the eye or mouth is closed. Figure 9(a) is a queue that stores the closure state of the mouth with M-CNN, and Figure 9(b) is a queue that stores the closure state of the eye with E-CNN.

We derive five features of eye and mouth from the queues, i.e., blinks, average blinking time, average blink time interval, PERCLOS, and FOM, as defined below:

(1) *Blinks.* The number of blinks is measured over a fixed time period. As the level of fatigue increases, the number would also change

(2) *Average Blinking Time (ABT).* It measures the average number of eye closures per blink in a fixed time period, which is often related to fatigue, calculated as

$$\mathrm{ABT} = \frac{n_{\mathrm{close}}}{N_{\mathrm{blinks}}}, \quad (3)$$

where n_{close} is the total number of eye-closed frames, and N_{blinks} is the total number of blinks over a period of time.

(3) *Average blink time interval (ABTI).* t refers to the average empty time intervals in a fixed time period, calculated as

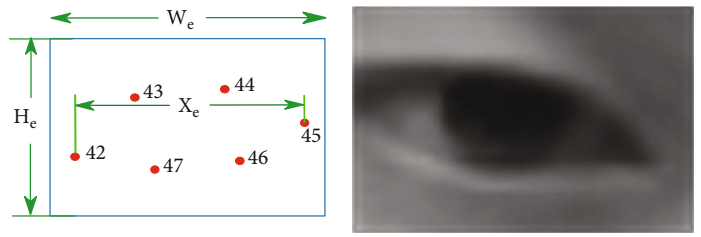
$$\mathrm{ABTI} = \frac{\sum_{i=n}^{n} t_{\mathrm{interval}}}{N_{\mathrm{blinks}} - 1}, \quad (4)$$

where n denotes the number of blink time intervals, and t_{interval} denotes the single blink time interval.

(4) *PERCLOS.* The ratio between the number of frames with closed eyes and the total number of frames in unit time, calculated as

$$\mathrm{PERCLOS} = \frac{n_{\mathrm{close}}}{N_{\mathrm{total}}}, \quad (5)$$

where n denotes the number of frames with closed eyes and N denotes the total number of frames.



(a) Eye feature points (b) Extracted eye area in an image

Figure 5: Eye identification.

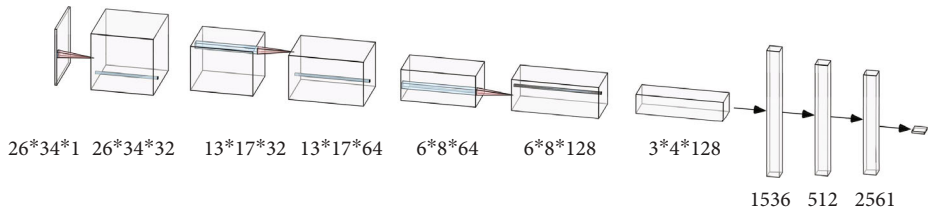
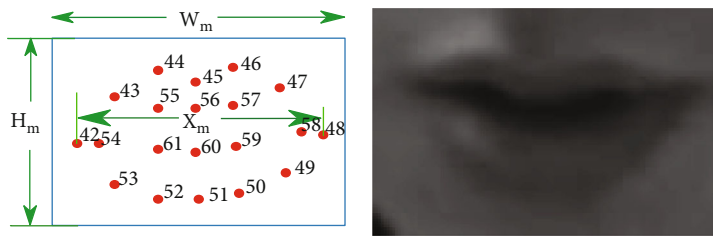


Figure 6: E-CNN structure.



(a) Mouth feature points (b) Extracted mouth area

Figure 7: Mouth identification.

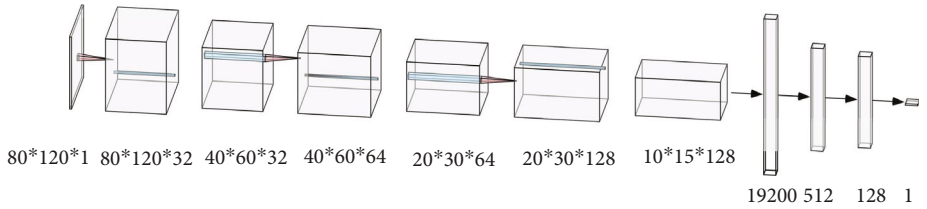
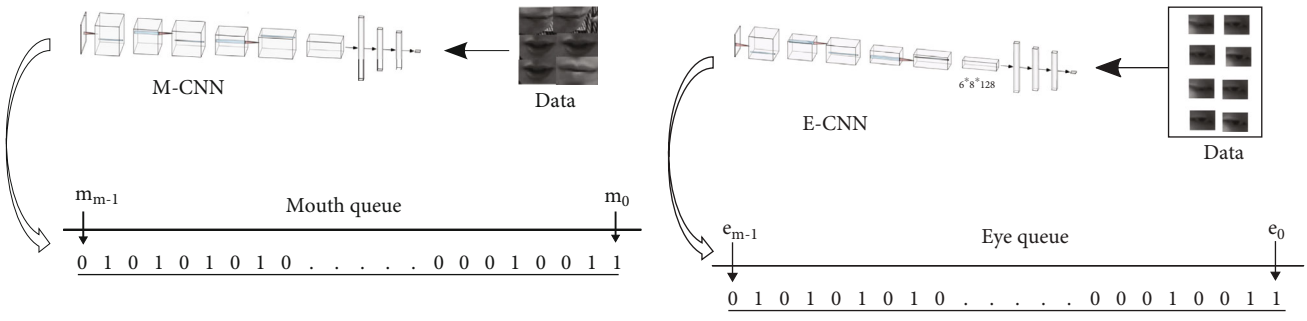


Figure 8: M-CNN structure.



(a) M-CNN detection results (b) E-CNN detection process

Figure 9: The process for obtaining the closure state of the mouth and eyes.

(5) *FOM*. Similar to PERCLOS, it refers to the ratio between the number of frames with closed mouth and the total number of frames in unit time, calculated as

$$\text{FOM} = \frac{n_{\text{closedmouth}}}{N_{\text{total}}}, \tag{6}$$

where $n_{\text{closedmouth}}$ denotes the number of frames with closed mouth and N_{total} denotes the total number of frames

3.2. Vocal Feature Extraction. It is critical to extract the most representative voice signal features for fatigue detection. In this paper, we employ MFCC feature extraction from voice signals as this unique cepstrum-based extraction method is more in line with the principle of human hearing, and it is also the most common and effective speech feature extraction algorithm. The MFCC extraction process is illustrated in Figure 10.

As shown in Figure 10, MFCC consists of seven steps, each of which has its own function and mathematical approach as discussed briefly below:

(1) *Preemphasis*. Preemphasis is a filtering method that emphasizes higher frequencies to balance the spectrum of voiced sounds that have a steep roll-off in the high-frequency region

(2) *Framing*. To facilitate speech analysis, voice signal can be divided into small segments, which are referred to as frames. Each frame contains N sampling points in an observation unit. Typically, N is set to be 256 or 512, and the time covered is about 20-30 ms

(3) *Windowing*. Voice is constantly changing in a long range and cannot be processed without fixed characteristics. Therefore, each frame is substituted into a window function, and the value outside the window is set to be 0. Commonly used window functions include square window, Hamming window, and Hanning window, etc. Considering the characteristics of a window function in the frequency domain, Hamming window is often used

(4) *Discrete Fourier Transform (DFT)*. Each windowed frame is converted into magnitude spectrum by applying DFT, calculated as

$$X(k) = \sum_{i=0}^{N-1} x(n) e^{(-j2\pi nk)/N}, \quad 0 \le K \le N-1, \tag{7}$$

where N is the number of points used to compute the DFT

(5) *Mel Spectrum*. Mel spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as Mel-filter bank. The Mel scale is approximately a linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz. The approximation of Mel from physical frequency is calculated as

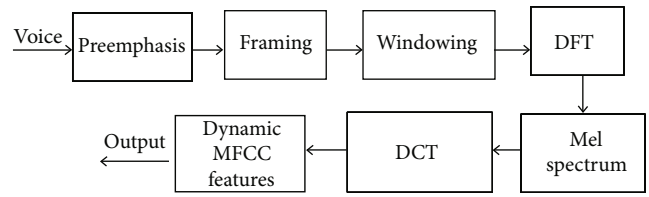


Figure 10: MFCC feature extraction.

$$f_{\text{Mel}} = 2569 \log_{10}\left(1 + \frac{f}{100}\right), \tag{8}$$

where f denotes the physical frequency in Hz, and f_{Mel} denotes the perceived frequency

(6) *Discrete Cosine Transform (DCT)*. DCT is applied to the transformed Mel frequency coefficients to produce a set of cepstral coefficients

(7) *Dynamic MFCC features*. Cepstral coefficients are usually referred to as static features, since they only contain information from a given frame. The extra information about the temporal dynamics of the signal is obtained by computing the first and second derivatives of cepstral coefficients

4. Experiments and Performance Evaluation

4.1. Experimental Platform. In this work, we use OpenCV [27] and Dlib libraries to process video dataset and use Keras and Sklearn frameworks to construct the model for fatigue detection. The entire detection system is implemented and tested on a Windows 10 PC equipped with 32GB of memory and a GPU with 8GB memory.

4.2. Dataset. For E-CNN, we collect 8,598 images with eyes open and 6,510 images with eyes closed. Altogether, we use 12,086 eye images for training and 3,022 images for testing. Similarly, for M-CNN, we collect 2,155 images with mouth open and 1,980 images with mouth closed. Altogether, we use 3,721 mouth images for training and 414 images for testing.

We also collect video and audio data of air traffic controllers in real operation. We collect 14,673 video and audio clips, where the length of each video is 15 seconds and the length of each audio is 7 seconds. Accordingly, we obtained 14,673 facial features and MFCC features from such video and audio data, out of which 11,738 are used for training the proposed FV-Stacking ensemble learning model, and 2,935 are used for testing.

4.3. Experiments. For the video data, we use OpenCV and Dlib to extract the eyes and mouth of each air traffic controller in each video frame, and then we use E-CNN and M-CNN models to identify the state of the eyes and mouth.

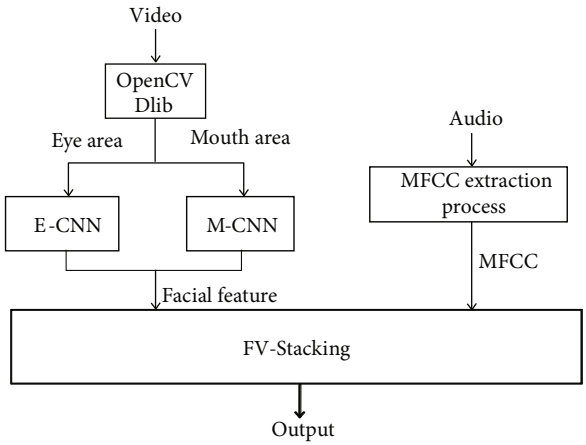


Figure 11: A schematic block diagram.

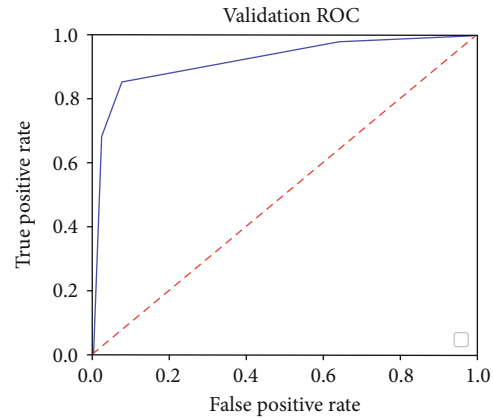


Figure 12: AUC based on ROC.

Table 1: the performance of CNN.

Model	Recall	Precision	Accuracy	f_1 score	AUC
E-CNN	98%	98%	98%	98%	0.99
M-CNN	97%	98%	97%	97%	0.99

Finally, five features are calculated, including blinks, average blinking time, average blink time interval, PERCLOS, and FOM. For the audio data, we obtained an MFCC feature vector of size 20×244 through the MFCC feature extraction process.

The facial features (i.e., blink times, average blinking time, average blink time interval, PERCLOS, and FOM) and MFCC features extracted from the audio are passed to the ensemble learning model as input. We use FV-Stacking to combine facial features and MFCC features to determine whether or not the air traffic controller is fatigued. The overall detection process is illustrated in Figure 11.

4.4. Experimental Results and Analysis. To verify the classification performance of M-CNN, ECNN, and FV-Stacking, we consider recall rate, precision, accuracy, f_1 score, and AUC (Area Under Curve) as the main performance metrics in our experiments, as defined in the following:

(1) Recall

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

where TP and FN denote the number of true positives and the number of false-negatives, respectively. This metric represents the proportion of positive samples that are correctly identified as a percentage of the total positive samples

(2) Precision

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{10}$$

where FP denotes the number of false-positives. This metric represents the portion of correctly identified positive samples as a percentage of all samples that are identified as positive

(3) Accuracy

$$\text{accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FP} + \text{FN}}, \tag{11}$$

where TN denotes the number of true negatives. This metric represents the proportion of correctly classified samples to the total number of samples

(4) f_1 score

$$f_1 = \frac{2 \cdot \text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} \tag{12}$$

This metric is based on the harmonic average of the recall rate and the precision rate.

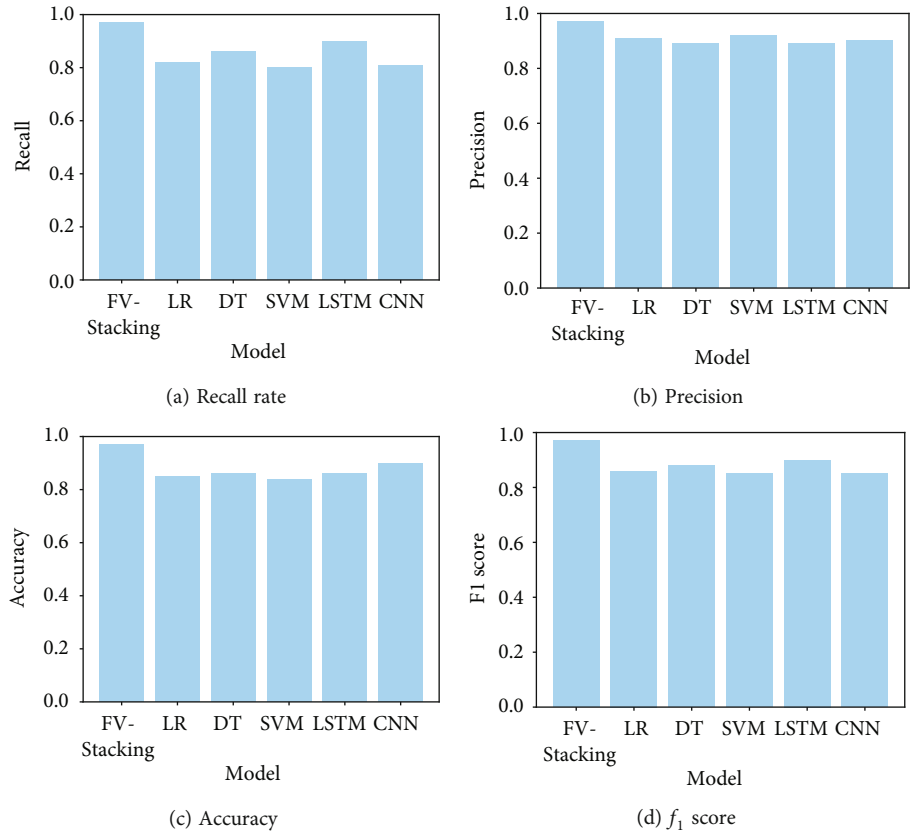
(5) *Area Under Curve (AUC).* A schematic diagram of the ROC (Receiver Operating Characteristic) curve is plotted in Figure 12. The horizontal axis of the curve is the false positive rate, calculated as

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}, \tag{13}$$

while the vertical axis is the true positive rate, calculated as

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{14}$$

In Figure 12, the area under the ROC curve and the horizontal axis is defined as the Area Under Curve (AUC). Obviously, the value of this area is no greater than 1. Moreover, because the ROC curve is generally above the line $y = x$,



(a) Recall rate

(b) Precision

(c) Accuracy

(d) f_1 score

(e) AUC

Figure 13: Performance comparison between different models.

Table 2: Performance comparison between different methods.

	Our method	Work by Zhang et al. [18]			Work by Nie et al. [19]			Work by Gu et al. [20]		
		LR	SVM	KNN	LR	SVM	KNN	LR	SVM	KNN
Precision	97%	85%	85%	88%	85%	86%	89%	0.83%	85%	73%
Accuracy	97%	84%	84%	88%	85%	85%	88%	82%	82%	71%
Recall	97%	86%	86%	89%	85%	85%	89%	83%	85%	70%
f_1 score	97%	85%	85%	88%	85%	85%	89%	83%	83%	70%
AUC	0.99	0.93	0.93	0.92	0.93	0.93	0.92	0.91	0.91	0.83

the value range of AUC is between 0.5 and 1. The closer the AUC is to 1.0, the better performance the detection method achieves.

The performance of E-CNN and M-CNN is shown in Table 1.

To evaluate the performance of FV-Stacking, we compare and analyze the recall rate, precision, accuracy, f_1 score, and AUC of single models and FV-Stacking. The results are plotted in Figure 13.

From Figure 13, we observe that the best recall rate of signal models is 90%, the precision is 92%, the accuracy is 90%, the f_1 score is 90%, and the AUC is 0.96. The recall of FV-Stacking proposed in this paper reaches 97%, the precision reaches 97%, the accuracy reaches 97%, the f_1 score reaches 97%, and the AUC reaches 0.99. These results show that FV-Stacking consistently outperforms any single model.

In some other methods, different features are used for fatigue detection. For example, in the work by Zhang et al. [18], fatigue is judged by PERCLOS and blinking frequency. In the work by Nie et al. [19], fatigue is judged by blink time, PERCLOS, blinks, and blink frequency. In the work by Gu et al. [20], fatigue is judged by PERCLOS and FOM. One common strategy is to determine the fatigue state by setting fixed thresholds for different characteristics. For example, in the work by Zhang et al. [18], the PERCLOS threshold is set to be 0.25, and in the work by Nie et al. [19], the PERCLOS threshold is set to be 0.06. In [20], Gu et al. set the PERCLOS threshold to be 0.5. However, in different scenarios, such fixed thresholds may not always yield the best performance. In order to mitigate the impact of thresholds on the performance, we combine different features using various machine learning models including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Logistic Regression (LR) to compare the performance of different fatigue detection methods, as summarized in Table 2.

From Table 2, we observe that the best recall rate of different fatigue detection methods is 89%, the accuracy is 88%, the f_1 score is 88%, the precision is 89%, and the AUC is 0.93. The recall rate of the proposed FV-Stacking method reaches 97%, the accuracy reaches 97%, the f_1 score reaches 97%, the precision reaches 97%, and the AUC reaches 0.99. These results show that FV-Stacking consistently outperforms other fatigue detection methods.

5. Conclusion and Future Direction

Civil aviation aircraft has become an indispensable tool for our daily travel. Route management at airports is becoming increasingly complicated as the airport size and the aircraft volume continue to grow. Such intensive work leads to fatigue of air traffic controllers, which is one of the major factors for accidents.

We focused on the fatigue detection problem for air traffic controllers. To improve detection accuracy, we combined facial features including blinks, average blink duration, average blink interval, PERCLOS, and yawn frequency as well as the MFCC characteristics of voice signal. We designed an ensemble learning method for fatigue detection and used real-life video and audio data for performance evaluation.

This research has resulted in the following findings:

(1) Both M-CNN and E-CNN are able to accurately identify the open and closed state of the mouth and eyes

(2) By strategically combining facial and speech features, the proposed ensemble learning model, FV-Stacking, is able to achieve consistently better detection performance in comparison with single models and other detection methods, in terms of various performance metrics

Our work provides a new perspective for the development of fatigue detection methods by combining facial features and vocal features. The proposed approach achieves a high fatigue detection rate and has a great potential to effectively avoid accidents caused by the fatigue of air traffic controllers.

There are many alternative vocal features in addition to MFCC. It is of our future interest to experiment with other vocal features such as Single Frequency Filtering Cepstral Coefficients (SFFCC) [28, 29] and Zero-Time Windowing Cepstral Coefficients [30]. Moreover, we plan to incorporate some other features that may also reflect the fatigue state of air traffic controllers, such as sitting posture [12].

Data Availability

The data used in this paper is not public. We have signed a confidentiality agreement with Southwest Air Traffic Management Bureau of Civil Aviation of China because these facial videos and voices were collected from the air traffic controllers performing tasks in the real environment in the Management Bureau of Civil Aviation.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Yi Hu and Zhuo Liu are the co-first authors with equal contributions to this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. U1833126, U2033203) and the Foundation for Safety and Capacity Development of Civil Aviation Administration of China (Grant No. ASSA2020/16).

References

[1] R. Sun, Z. Yuan, L. Sun, and Y. Ma, "Analysis of safety trend in civil aviation of China," in *2017 4th International Conference on Transportation Information and Safety (ICTIS)*, pp. 852–857, Banff, AB, Canada, 2017.

[2] P. He and R. Sun, "Research on crosscorrelation, co-integration, and causality relationship between civil aviation incident

and airline capacity in China," *Sustainability*, vol. 14, no. 9, p. 4999, 2022.

[3] J. H. Marcus and M. R. Rosekind, "Fatigue in transportation: NTSB investigations and safety recommendations," *Injury Prevention*, vol. 23, no. 4, pp. 232–238, 2017.

[4] S. Lee and J. K. Kim, "Factors contributing to the risk of airline pilot fatigue," *Journal of Air Transport Management*, vol. 67, no. MAR., pp. 197–207, 2018.

[5] K. Jiang, F. Ling, Z. Feng, K. Wang, and C. Shao, "Why do drivers continue driving while fatigued? An application of the theory of planned behaviour," *Transportation Research Part A: Policy and Practice*, vol. 98, pp. 141–149, 2017.

[6] H. Wang, A. Dragomir, N. I. Abbasi, J. Li, N. V. Thakor, and A. Bezerianos, "A novel real-time driving fatigue detection system based on wireless dry EEG," *Cognitive Neurodynamics*, vol. 12, no. 4, pp. 365–376, 2018.

[7] S. Murugan, J. Selvaraj, and A. Sahayadhas, "Detection and analysis: driver state with electrocardiogram (ECG)," *Physical and Engineering Sciences in Medicine*, vol. 43, no. 2, pp. 525–537, 2020.

[8] A. Y. Awad and S. Mohan, "Real-time v2v communication with a machine learning-based system for detecting drowsiness of drivers," *International Journal of Interdisciplinary Telecommunications and Networking (IJITN)*, vol. 13, no. 4, pp. 35–50, 2021.

[9] A. Kuwahara, K. Nishikawa, R. Hirakawa, H. Kawano, and Y. Nakatoh, "Eye fatigue estimation using blink detection based on Eye Aspect Ratio Mapping(EARM)," *Cognitive Robotics*, vol. 2, pp. 50–59, 2022.

[10] L. W. Ko, O. Komarov, W. K. Lai, W. G. Liang, and T. P. Jung, "Eyeblink recognition improves fatigue prediction from single-channel forehead EEG in a realistic sustained attention task," *Journal of Neural Engineering*, vol. 17, no. 3, article 036015, 2020.

[11] H. G. Canada, M. J. O. Jaurigue, K. K. M. C. Antonio, J. B. Ilagan, and Y. T. Prasetyo, "Assessment of fatigue through working environment and hazardous driving behavior among drivers," in *2021 3rd International Conference on Management Science and Industrial Engineering*, pp. 158–166, Osaka, Japan, 2021.

[12] A. Verma, A. Goyal, and D. Kaur, "Fatigue detection," 2019, https://arxiv.org/abs/1911.10629.

[13] A. M. Williamson, A. M. Feyer, R. P. Mattick, R. Friswell, and S. Finlay-Brown, "Developing measures of fatigue using an alcohol comparison to validate the effects of fatigue on performance," *Accident Analysis & Prevention*, vol. 33, no. 3, pp. 313–326, 2001.

[14] J. De Vries, H. J. Michielsen, and G. L. Van Heck, "Assessment of fatigue among working people: a comparison of six questionnaires," *Occupational and Environmental Medicine*, vol. 60, no. >90001, pp. i10–i15, 2003.

[15] S. Arnau, T. Möckel, G. Rinkenauer, and E. Wascher, "The interconnection of mental fatigue and aging: an EEG study," *International Journal of Psychophysiology*, vol. 117, pp. 17–25, 2017.

[16] M. L. Chen, S. Y. Lu, and I. F. Mao, "Subjective symptoms and physiological measures of fatigue in air traffic controllers," *International Journal of Industrial Ergonomics*, vol. 70, pp. 1–8, 2019.

[17] D. F. Dinges and R. Grace, *PERCLOS: a valid psychophysiological measure of alertness as assessed by psychomotor vigilance*, Federal Highway Administration. Office of Motor Carriers. Brief, Washington, 1998.

[18] F. Zhang, S. Jingjing, L. Geng, and Z. Xiao, "Driver fatigue detection based on eye state recognition," in *2017 International Conference on Machine Vision and Information Technology (CMVIT)*, pp. 105–110, Singapore, 2017.

[19] B. Nie, X. Huang, Y. Chen, A. Li, R. Zhang, and J. Huang, "Experimental study on visual detection for fatigue of fixed-position staff," *Applied Ergonomics*, vol. 65, pp. 1–11, 2017.

[20] 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 Processing*, vol. 12, no. 12, pp. 2319–2329, 2018.

[21] Z. Shen, G. Pan, and Y. Yan, "A high-precision fatigue detecting method for air traffic controllers based on revised fractal dimension feature," *Mathematical Problems in Engineering*, vol. 2020, Article ID 4563962, 13 pages, 2020.

[22] O. Sagi and L. Rokach, "Ensemble learning: a survey. *Wiley interdisciplinary reviews*," *Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1249, 2018.

[23] H. Xia and C. Li, "Face recognition and application of film and television actors based on Dlib," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 1–6, Suzhou, China, 2020.

[24] A. F. Iaquinta, A. C. D. S. Silva, A. F. Júnior, J. M. de Toledo, and G. V. von Atzingen, "EEG multipurpose eye blink detector using convolutional neural network," 2021, https://arxiv.org/abs/2107.14235.

[25] L. O. Chua, "CNN: a vision of complexity," *International Journal of Bifurcation and Chaos*, vol. 7, no. 10, pp. 2219–2425, 1997.

[26] R. M. Salman, M. Rashid, R. Roy, M. M. Ahsan, and Z. Siddique, "Driver drowsiness detection using ensemble convolutional neural networks on YawDD," 2021, https://arxiv.org/abs/2112.10298.

[27] M. Beyeler, *Machine learning for OpenCV: a practical introduction to the world of machine learning and image processing using OpenCV and python*, Packt Publishing Ltd, 2017.

[28] S. R. Kadiri and B. Yegnanarayana, *Analysis and Detection of Phonation Modes in Singing Voice Using Excitation Source Features and Single Frequency Filtering Cepstral Coefficients (SFFCC)*, Interspeech, 2018.

[29] S. R. Kadiri, R. Kethireddy, and P. Alku, *Parkinson's Disease Detection from Speech Using Single Frequency Filtering Cepstral Coefficients*, Interspeech, 2020.

[30] R. Kethireddy, S. R. Kadiri, S. Kesiraju, and S. V. Gangashetty, *Zero-Time Windowing Cepstral Coefficients for Dialect Classification*, Aalto University, 2020.