aerospace

MDPI

Article

Air Traffic Controller Fatigue Detection by Applying a Dual-Stream Convolutional Neural Network to the Fusion of Radiotelephony and Facial Data

Lin Xu [1], Shanxiu Ma [2], Zhiyuan Shen [2,*] and Ying Nan [1]

[1] College of Astronautics, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; xulin19851116@163.com (L.X.); nanying@nuaa.edu.cn (Y.N.)
[2] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China; masxiu@nuaa.edu.cn
* Correspondence: shenzy@nuaa.edu.cn

Abstract: The role of air traffic controllers is to direct and manage highly dynamic flights. Their work requires both efficiency and accuracy. Previous studies have shown that fatigue in air traffic controllers can impair their work ability and even threaten flight safety, which makes it necessary to carry out research into how to optimally detect fatigue in controllers. Compared with single-modality fatigue detection methods, multi-modal detection methods can fully utilize the complementarity between diverse types of information. Considering the negative impacts of contact-based fatigue detection methods on the work performed by air traffic controllers, this paper proposes a novel AF dual-stream convolutional neural network (CNN) architecture that simultaneously extracts controller radio telephony fatigue features and facial fatigue features and performs two-class feature-fusion discrimination. This study designed two independent convolutional processes for facial images and radio telephony data and performed feature-level fusion of the extracted radio telephony and facial image features in the fully connected layer, with the fused features transmitted to the classifier for fatigue state discrimination. The experimental results show that the detection accuracy of radio telephony features under a single modality was 62.88%, the detection accuracy of facial images was 96.0%, and the detection accuracy of the proposed AF dual-stream CNN network architecture reached 98.03% and also converged faster. In summary, a dual-stream network architecture based on facial data and radio telephony data is proposed for fatigue detection that is faster and more accurate than the other methods assessed in this study.

Keywords: human factor; fatigue detection; dual-stream network; radio telephony; facial image

Citation: Xu, L.; Ma, S.; Shen, Z.; Nan, Y. Air Traffic Controller Fatigue Detection by Applying a Dual-Stream Convolutional Neural Network to the Fusion of Radiotelephony and Facial Data. *Aerospace* **2024**, *11*, 164. https://doi.org/10.3390/aerospace11020164

Academic Editor: Julius Keller

Received: 25 December 2023
Revised: 9 February 2024
Accepted: 14 February 2024
Published: 17 February 2024

1. Introduction

The subsiding of the COVID-19 epidemic and the associated ongoing increase in the number of flights is further increasing the workload of air traffic controllers, thereby increasing the problem of controller fatigue [1]. At busy airports, controllers who continuously issue control instructions are likely to experience fatigue symptoms such as dry mouth and difficulty speaking [2]. Studies have shown that fatigue can significantly reduce the reaction speed, judgment accuracy, and decision-making ability of controllers. Kelly D. [3] studied the human factors in some aviation accidents from 2007 to 2017 and found that fatigue is an important cause of accidents. Abd-Elfattah H. M.'s research shows that fatigue has a negative impact on human perception and decision making [4]. Fatigue may cause errors, omissions, and forgetfulness in the work of controllers, thereby threatening the safe operation of flights [5]. In September 2011, a controller in Japan fell asleep while on duty in the early morning, causing an incoming plane to lose contact with the ground for more than 10 min. Civil aviation safety incidents have occurred due to controller fatigue

resulting in sleeping on duty, indicating that controller fatigue has always been a potential threat to the efficient and safe operation of civil aviation.

Interventions to effectively curb the negative impacts of fatigue require research into controller fatigue detection. Fatigue detection methods are usually divided into subjective and objective methods: subjective methods usually involve questionnaires, while objective methods are based on objective data such as physiological indicators of the subjects. Objective detection methods can also be divided into contact-based and non-contact-based methods according to whether the detection equipment makes physical contact with the subject. Based on detection indicators, fatigue detection methods can also be divided into single-modality detection methods based on a single data source such as audio or facial data and multi-modal detection methods based on multiple data sources.

Contact-based fatigue detection equipment can interfere with the work of controllers, which has prompted many researchers to investigate non-contact-based fatigue detection methods. Researchers have found that fatigue is associated with numerous facial features such as eye closure rate, eyelid distance, percentage of eye open [6], blinking frequency, mouth breathing [7,8], and other facial features [9,10]. Liang [11] analyzed eye features when controllers were working and proposed a deep-fusion neural network for eye position and eye state detection. Deng et al. [12] studied the relationship between the percentage of the pupil covered by the eyelid over time and the fatigue state of controllers. Li K. [13] focused on analyzing the information of the eyes and mouth and fused the fatigue information from different facial regions through multi-source fusion, proposing an accurate recognition algorithm called Recognizing the Drowsy Expression (REDE). Zhang et al. [14] revealed the fatigue state of controllers based on changes in pupil size. These studies have shown that fatigue information is indeed available from facial features. However, fatigue is not only reflected in the eyes, so research focusing only on the eye area ignores other information available from the face.

Many studies on the facial fatigue features of car drivers have shown that actions such as mouth breathing contain significant fatigue information [15]. Devi et al. [16] proposed a fatigue state discrimination method based on a fuzzy inference system to perform car driver fatigue state discrimination by fusing mouth breathing and the fatigue state of the eyes. Li et al. [17] improved the Tiny YOLOV3 convolutional neural network and evaluated the fatigue state of car drivers based on both eye and mouth features.

In contrast to car drivers, air traffic controllers need to speak as an integral part of their workflow, which provides information for fatigue state detection. Moreover, collecting radio telephony data has little impact on the work of controllers, making such data ideal for controller fatigue detection. Audio features such as hesitations, silent pauses, prolongation of final syllables, and the syllable articulation rate in a fatigue state differ significantly from those in a normal state [18], thus confirming the possibility of analyzing controller fatigue state through audio analysis. Wu [19] proposed an audio fatigue detection algorithm based on traditional Mel-Frequency Cepstral Coefficients (MFCC) and added an adaptive mechanism to the algorithm to successfully classify the fatigue audio of air traffic controllers. Shen and others [20] revealed significant differences in the fractal dimensions of radio telephony under different fatigue states. He [21] analyzed the radio telephony of controllers under different fatigue states and found that features such as the audio rate and pitch can be utilized by a k-means++ algorithm to classify the fatigue state. Shen [22] applied a densely connected convolutional autoencoder to neural networks to classify fatigue radio telephony.

In addition, there are fatigue detection studies based on other data sources. We have summarized the fatigue detection methods according to the data sources of fatigue detection in Table 1.

Table 1. Summary of fatigue detection methods and characteristics.

Category	Method		Principle	Accuracy	Usability
Subjective Detection Methods	Subjective Feeling Rating Method		Determine fatigue level based on subjective fatigue feeling.	Medium	Low
	Fatigue Rating Scale Method		Design scales to rate fatigue level based on fatigue characterization indicators.	Medium	Medium
Objective Detection Methods	Contact Type	Electroencephalogram Measurement Method	Different brain wave frequencies when the cerebral cortex is in different states.	High	Low
		Electrocardiogram Measurement Method	Heart rate time–frequency domain indicators are significantly related to the degree of fatigue.	High	Low
		Electromyogram Measurement Method	Monitor the bioelectric changes when muscle cells are active.	High	Low
		Dynamic Heart Rate Method	There is a close relationship between heart rate and muscle fatigue when engaging in physical operations.	High	Medium
	Non-contact Type	Facial State Recognition Method	Detect fatigue by analyzing and recognizing facial features.	Medium	High
		Voice Frequency Analysis Method	Voice features change under fatigue state.	Medium	Medium

In comparison, multi-modal fatigue state detection data contain richer and more-detailed fatigue information since they are affected by multiple aspects of the fatigue state. The processing of weights for different types of modal information is a major difficulty when fusing multi-modal information. In order to dynamically adjust the degree of influence of two types of features on the detection results, the fusion of audio features and facial features can be weighted separately as $\delta_{\text{sum}} = \vartheta\delta_v + (1-\vartheta)\delta_a$ and weighted product $\delta_{\text{prod}} = \delta_v^{\vartheta}\delta_a^{1-\vartheta}$, where the weight factor $\vartheta \in [0, 1]$ and δ_v and δ_a are facial features and audio features, respectively [23]. Authors have adjusted the weight factors through experiments to achieve the optimal fusion of audio features and facial features and used δ_{sum} and δ_{prod} as the fusion features. However, manual adjustments are both time-consuming and inefficient.

In order to solve the problem of accurately detecting the fatigue state of controllers without affecting their work, this paper proposes a multi-modal fatigue feature detection network for controllers based on audio data and facial data. The proposed network applies two independent convolution processes to audio data and facial data, fuses audio features with facial features at the feature level, and finally realizes fatigue state discrimination using the Softmax function. Compared with the weighted feature-fusion method, the feedforward neural network can automatically correct the weights of each neuron during backpropagation, not only by dynamically adjusting the weights between different modalities of features but also between different features within the same modality and between different elements within the same feature. This approach can reveal key information in multi-modal features and improve the accuracy of fatigue state discrimination.

This paper was organized as follows: Section 2 introduces the proposed AF dual-stream CNN architecture, explaining the processing of audio data and facial data as well as the fusion and discrimination steps of the two features. Section 3 introduces comparative experiments of the AF dual-stream CNN to verify the effectiveness of the network. Finally, conclusions are drawn and future work is discussed in Section 4.

2. AF Dual-Stream CNN: A Dual-Stream CNN for Audio and Facial Images

This section first introduces the extraction of audio features in the dual-stream network and then introduces the proposed AF dual-stream CNN architecture.

2.1. Audio Feature Extraction

Various vocal feature extraction methods have different focuses on the features they extract. Therefore, the simultaneous use of multiple vocal feature extraction methods can

comprehensively reflect the differences before and after vocal fatigue. This study selected five commonly used audio features from five perspectives as audio fatigue features [24], as briefly introduced below.

(a) Zero-crossing rate

In the waveform diagram, the zero-crossing rate (ZCR) [25] represents the number of times the waveform crosses the *X*-axis. The short-time zero-crossing rate of an audio single A_i is calculated as

$$Z_n(A_i) = \sum_{m=-\infty}^{\infty} \left| \mathrm{sgn}[A_i(m)] - \mathrm{sgn}[A_i(m-1)] \right| w(m) \tag{1}$$

where $\mathrm{sgn}(n)$ is the following sign function:

$$\mathrm{sgn}[A_i(n)] = \begin{cases} 1 & A_i(n) \geq 0 \\ -1 & A_i(n) < 0 \end{cases} \tag{2}$$

The $w(n)$ function is

$$w(n) = \begin{cases} 1/2N & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This feature can be used to represent spectral and noise changes.

(b) Chromagram

Chromaticity features [26] are collectively referred to as the chroma vector and the chromagram. The chromaticity vector is a vector containing 12 elements, which represent the energy in 12 levels within a period of time. Energy of the same level in different octaves is accumulated. The chromaticity diagram is a sequence of chromaticity vectors. The chroma features of an audio sample A_i are represented as

$$C(A_i) = C(FFT(A_i)) \tag{4}$$

where FFT is the fast Fourier transform.

$C(A_i)$ can capture harmonic information in an audio signal and has high robustness.

(c) Mel-frequency cepstral coefficients

MFCC [27] are a widely used cepstral parameter extracted in the mel-scale frequency domain:

$$Mf(A_i) = DCT(Mel(FFT(P_{re}(A_i)))) \tag{5}$$

The MFCC feature of an audio signal A_i can be obtained by applying preprocessing, the fast Fourier transform, triangular mel-filter processing, and the discrete Fourier transform. This feature conforms to the auditory characteristics of human ears and can convert raw audio into separable and recognizable feature vectors.

(d) Root mean square

The size of a frame of signals can be quantified as its root mean square (*RMS*) value [28], which is essentially a set of arithmetic mean values:

$$RMS(A_i) = \sqrt{\frac{1}{K} \cdot \sum_{k=t \cdot K}^{(t+1) \cdot K-1} s(k)^2} \tag{6}$$

where $\frac{1}{K} \cdot \sum_{k=t \cdot K}^{(t+1) \cdot K-1} s(k)^2$ represents the average energy at all sampling points in frame t. $RMS(A_i)$ has the advantage of not being sensitive to outliers.

(e) Mel spectrogram

The mel-spectrogram [29] is calculated by mapping the power spectrum onto the mel frequency scale. This can capture the spectral information of audio signals and reflect changes therein over time. The mel-spectrogram of an audio signal A_i is

$$M_s(A_i) = \text{MelSpectrogram}(A_i) \tag{7}$$

We concatenate the five types of audio features of an audio signal to obtain its feature matrix A_{if}:

$$A_{if} = \begin{bmatrix} Z_n(A_i) \\ C(A_i) \\ Mf(A_i) \\ RMS(A_i) \\ M_s(A_i) \end{bmatrix} \tag{8}$$

We then input the combined audio features A_{if} of each audio A_i into the designed audio data stream network.

In Figure 1, blue, yellow, and green, respectively, represent the distribution of voice feature values when "Notfatigue", "Mildfatigue", and "Fatigue" are present, and we can see that the voice features of the three fatigue states are intertwined with each other, so the classification of fatigue speech is difficult.

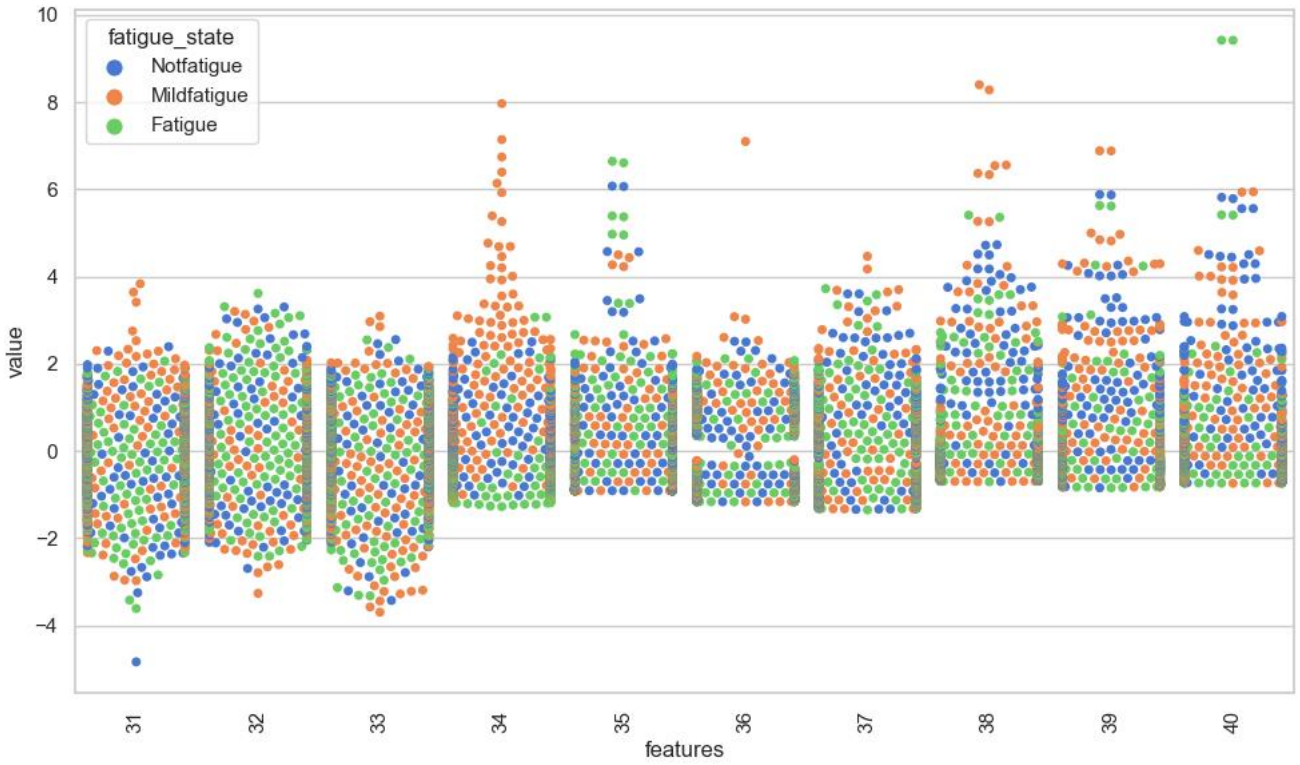

Figure 1. A 30–40 voice feature clustering scatter plot.

2.2. AF Dual-Stream CNN

In order to accurately and rapidly detect a controller's fatigue state, the characteristics of the controller's work [30] are utilized in this paper to propose a fatigue detection network AF dual-stream CNN based on audio data and facial data. The network architecture includes an audio convolution module, a facial convolution module, and a fully connected layer.

(a) Audio data stream: convolution module based on audio data

In Section 2.1, we introduced the five types of features of audio. In our model, we pass audio feature A_{if} to a one-dimensional audio convolution module that includes four

convolution layers and four pooling layers. The audio features are used as audio fatigue features after convolution processing. The essence of the audio convolution module is performing function mapping with A_{if} as an independent variable so that the convolution operation of the audio data stream can be recorded as function $ADS\left(A_{if}\right)$, and the convolution operation process of the audio data stream is

$$A_{ff} = ADS\left(A_{if}\right) \tag{9}$$

where audio fatigue feature A_{ff} is produced by the convolution processing of the audio features.

(b) Facial data stream: convolution module based on facial data

We designed a two-dimensional convolution module for processing facial features. This convolution module first performs three convolution operations, followed by pooling operations, two further convolution operations, and, finally, more pooling operations. Suppose a facial picture with a resolution of $n \times n$ P_i is denoted as $P_{n\times n}$:

$$P_{n\times n} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,n} \end{bmatrix} \tag{10}$$

where $x_{i,j} \in (0, 255)$ and $i, j \in (1, n)$; $x_{i,j}$ represents the value of a pixel in a facial image.

This means that the independent variable in the convolution operation of the facial data stream is $P_{n\times n}$, and the convolution mapping of the facial data stream can be written as

$$F_{ff} = FDS(P_{n\times n}) \tag{11}$$

where F_{ff} is an $m\times m$ matrix representing the facial fatigue features after convolution processing of the facial image. In our network, n = 48 and m = 12.

(c) Feature fusion and fatigue state discrimination

The processing results for the audio and facial data streams are fused in the fully connected layer, and, finally, the fused features are input into the Softmax classifier for classifying three fatigue states.

For a facial image output $F_{ff} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,m} \end{bmatrix}$ after the facial convolution data stream, we extend it to $F_{ff}{}' = \begin{bmatrix} x_{1,1} \\ \vdots \\ x_{m,m} \end{bmatrix}$ and then perform feature concatenation in the input layer of the fully connected layer. The feature concatenation process is as follows:

$$F_f = \begin{bmatrix} A_{ff} \\ F_{ff}{}' \end{bmatrix} = \begin{bmatrix} Z_n(A_i) \\ C(A_i) \\ Mf(A_i) \\ RMS(A_i) \\ M_s(A_i) \\ x_{1,1} \\ \vdots \\ x_{12,12} \end{bmatrix} \tag{12}$$

We then input fatigue feature FF_f containing audio information and facial information into the trained fully connected layer for feature fusion. The fully connected layer as a function FCL$\left(F_f\right)$ results in the process of feature fusion and detection being described as

$$FF = \text{FCL}\left(F_f\right) \tag{13}$$

$$\text{result} = \text{Solftmax}(FF) \tag{14}$$

where result $\in \{NoFatigue,\ MildFatigue,\ Fatigue\}$. The detailed network architecture is shown in Figure 2. The detailed algorithm process is shown in Algorithm 1.

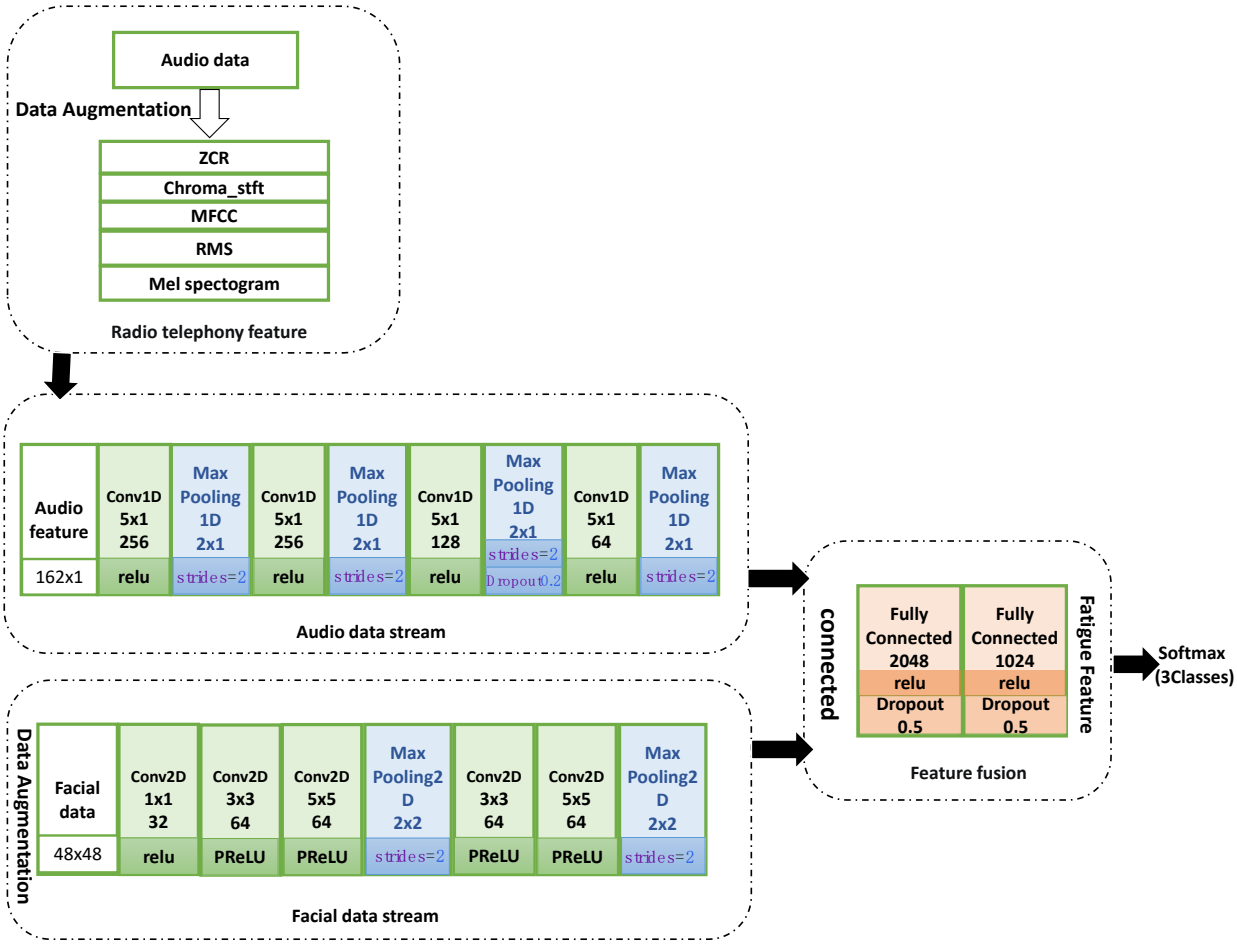

Figure 2. Network structure of the AF dual-stream CNN that includes three modules: audio data stream, facial data stream, and feature fusion.

Algorithm 1: AF dual-stream CNN

Input: A_i, P_i
Output: result
Initialize: initialize F_f
Step 1: initialize A_{ff} and $F_{ff}^{'}$
For A_i, initialize $Z_n(A_i)$, $C(A_i)$, $Mf(A_i)$, $RMS(A_i)$, and $M_s(A_i)$ and, by using Equations (1) and (4)–(7), initialize A_{if}.

$$A_{if} = \begin{bmatrix} Z_n(A_i) \\ C(A_i) \\ Mf(A_i) \\ RMS(A_i) \\ M_s(A_i) \end{bmatrix}$$

Then, $A_{ff} = ADS\left(A_{if}\right)$ according to Equation (9).

Algorithm 1: *Cont.*

For $\boldsymbol{P_i}$, initialize $\boldsymbol{F_{ff}} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,m} \end{bmatrix}$ according to Equation (11), $\boldsymbol{F_{ff}}' = \begin{bmatrix} x_{1,1} \\ \vdots \\ x_{m,m} \end{bmatrix}$

Step 2: initialize $\boldsymbol{F_f}$

$$\boldsymbol{F_f} = \begin{bmatrix} \boldsymbol{A_{ff}} \\ \boldsymbol{F_{ff}}' \end{bmatrix}$$

Step 1: Fully connected layer update

For $E = 1$ to the number of iterations,
train the fully connected layer using $\boldsymbol{F_f}$ of each $\boldsymbol{A_j}$ and $\boldsymbol{P_j}$ in the training set;
according to the difference between the input and output labels, update the parameters in the fully connected layer using a backpropagation algorithm;
$E = E + 1$.
End

Step 2: Fatigue state discrimination

Initialize $\boldsymbol{FF}$ of $\boldsymbol{A_i}$ and $\boldsymbol{P_i}$ according to Equation (13).
Output the fatigue label according to Equation (14): **result** = **Solftmax**($\boldsymbol{FF}$)

3. Fatigue Detection Experiments

This section introduces the experimental environment, including the experimental data and parameter settings of the proposed network used in the experiments. The comparative experiments performed with other mainstream methods are also described.

3.1. Experimental Setup

In collaboration with the Jiangsu Air Traffic Control Bureau, we collected radio telephony data and facial data from 14 certified air traffic controllers while they were working. We used the same aircraft model as the Jiangsu Air Traffic Control Bureau to ensure that the simulation environment was identical to the actual working environment. The data collection experiments were conducted on a control simulator produced by China Electronics Technology Group Corporation. The control simulator system was divided into two parts: the tower seat and the captain's seat. The experimental environment of the control tower seat is shown in Figure 3. The subjects were located in the tower seat, and the equipment used included apron display screens, electronic progress sheet display screens, control call recording equipment, and facial data collection cameras. The subjects comprised seven males and seven females, all of whom were licensed tower controllers in the Air Traffic Management Bureau and came from East China; hence, their radio telephony had similar pronunciation characteristics. They had between 3 and 8 years of work experience, and the workload of the air traffic controllers in the simulation scenario was similar to the workload in their actual working environment. All the subjects had rested sufficiently (>7 h) for the two nights before the experiments and were prohibited from consuming food or alcohol that might affect the experimental results. All the experimenters were fully familiar with the control simulator system. All the subjects were informed of the experimental content and had the right to withdraw from participating in the experiments at any time.

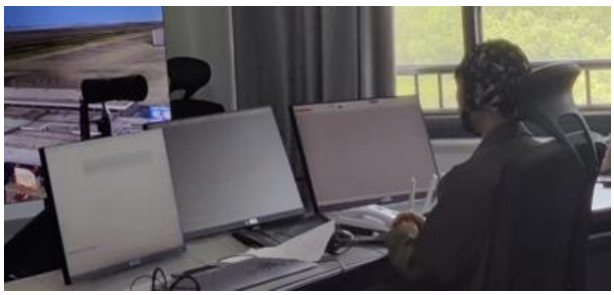

Figure 3. Fatigue data collection experiment subjects (tower seat).

The experiments were conducted at 14:00 every day from 4 April to 27 April 2023. During this period, one controller was assigned to conduct experiments in the control seat every day. During each experimental day, we asked the subjects to complete six sets of control simulation tasks, each set of which had a volume of 20 flights and lasted for 30 min, and with radio telephony and facial data only recorded while they were in the control seat during the experiments. After each set of tasks had been completed, we allowed the subjects to rest for 5 min and complete the 9-point Karolinska Sleepiness Scale during this rest period to assess their fatigue status as a score from 1 to 9 [31]. The result was used to categorize the controller's fatigue status as no fatigue, mild fatigue, or severe fatigue [32]. At the beginning of the experiment, the scale assessment results showed that the subjects were mostly in a no-fatigue state. After the experiment had been conducted for a while, the assessment results of some scales showed that the subjects appeared to be fatigued. The typical fatigue features are shown in Figure 4.

Figure 4. Facial schematic diagram of the controller in a fatigued state (**left image**) and in a non-fatigued state (**right image**).

After the experiments, we preprocessed the facial and radio telephony data. After processing, each radio telephony sample lasted about 2 s and was associated with a facial image, with both of them being used as a set of data. Finally, 1602 facial images were obtained, corresponding to 1602 radio telephony samples for the same period. The data of all the subjects were processed, which finally yielded 496, 543, and 563 sets of data for the no-fatigue, mild-fatigue, and severe-fatigue states, respectively. The average age of the participants was 29.2, with a standard deviation of 2.9. The data distribution is presented in Table 2.

Table 2. Data distribution.

Subject Number	Gender	Age, Years	No Fatigue	Mild Fatigue	Severe Fatigue	Total
1	Male	30	35	39	40	109
2	Male	28	30	40	41	116
3	Male	27	34	38	39	111
4	Male	28	35	39	40	109
5	Male	30	42	46	47	127
6	Male	29	30	34	35	104
7	Male	32	32	36	37	105
8	Female	35	33	37	38	108
9	Female	28	37	41	42	120
10	Female	29	38	42	43	123
11	Female	31	39	43	44	121
12	Female	30	40	44	45	118
13	Female	29	36	40	41	117
14	Female	29	35	39	40	114

Our dual-stream network was trained using each set of data as an input sample, and each set of data participating in the network training had undergone data-enhancement processing. For the radio telephony data, we injected noise and performed slice processing as well as compression and expansion processing in the time domain. For the facial images, we used an iterator to introduce random disturbances into the data for each round of training, including scaling and slight rotation of the images. Using these preprocessing methods helped us to increase the generalization of the model.

The experiments were conducted using Python 3.6 in the Windows operating system. To ensure the reliability of the experimental results, the experiments were repeated multiple times, from which average values were determined.

The framework of the proposed dual-stream network used in the experiments is shown in Figure 1, and the detailed parameter settings are listed in Table 3.

Table 3. Parameter settings for the audio data stream. The same padding method was applied in all cases.

Network Layer	Number of Kernels	Kernel Size	Stride	Dropout	Activation Function	Output Size
Audio feature						162 × 1
Conv1D	256	5	1	0	Relu	162 × 256
MaxPooling1D	0	2	2	0		81 × 256
Conv1D	256	5	1	0	Relu	81 × 256
MaxPooling1D	0	2	2	0		41 × 256
Conv1D	128	5	1	0	Relu	41 × 128
MaxPooling1D	0	2	2	0.2		21 × 128
Conv1D	64	5	1	0	Relu	21 × 64
MaxPooling1D	0	2	2	0		11 × 64

The parameter settings of the facial data stream are as shown in Table 4.

The features obtained from the facial data stream and the voice data stream are trained in the fully connected layer to achieve fatigue classification. The parameter settings of the fully connected layer are as shown in Table 5.

During the experiments, in order to facilitate the display of the detection results for a single modality, we used the network formed by connecting Tables 3 and 5 as the audio data stream model. The network formed by connecting Tables 4 and 5 as the facial data stream model. We subsequently conducted experiments on single-modality and multi-modal networks.

Table 4. Parameter settings for the facial data stream. The same padding method was applied in all cases.

Network Layer	Number of Kernels	Kernel Size	Stride	Dropout	Output Size
Facial data					48 × 48 × 1
Conv2D	32	1 × 1	1	Relu	48 × 48 × 32
Conv2D	64	3 × 3	1	Prelu	48 × 48 × 64
Conv2D	64	5 × 5	1	Prelu	48 × 48 × 64
MaxPooling2D	0	2 × 2	2		24 × 24 × 64
Conv2D	64	3 × 3	1	Prelu	24 × 24 × 64
Conv2D	64	5 × 5	1	Prelu	24 × 24 × 64
MaxPooling1D	0	2 × 2	2		12 × 12 × 64

Table 5. Parameter settings for the fully connected layer.

Network Layer	Input	Output	Activation Function	Dropout	Classifier
Fully connected 2048	9920	2048	Relu	0.5	None
Fully connected 1024	2048	1024	Relu	0.5	Softmax

3.2. Experimental Results

For the radio telephony data, we selected commonly used and classic methods to classify the audio features. In the experiments, when we only used the audio data stream to train the fully connected layer of the neural network, the accuracy of the audio data stream model reached 62.88%, while the AF dual-stream CNN detection accuracy based on the radio telephony data and the facial data reached 98.03%. In the experiments, we aimed to set other model parameters to their optimal values in order to ensure that all comparisons were unbiased. The parameter settings for the other models were as follows:

a. For the SVC model, the penalty coefficient C was 10, the kernel function used the radial basis function, and the randomness was set to 69.
b. For the KNN model, the number of neighbors was set to five, the prediction weight function was inversely proportional to the distance, and the brute force algorithm was used. The leaf size passed to the nearest-neighbor search algorithm was 30.
c. For the random forest model, the number of trees was set to 500, the random state was 69, and the maximum number of features was the square root of the number of sample features. The node split criterion was the information gain entropy.
d. For the multilayer perceptron classifier unscaled MLP model, randomness was set to 69, and data scaling was not performed during testing.
e. For the multilayer perceptron classifier standard scaled MLP model, randomness was set to 69, and data scaling was performed during testing.

The experimental results are presented in Table 6.

Table 6. Comparison of audio detection model accuracies for different models.

Model Name	Accuracy
SVC model (C = 10)	51.40%
KNN model (K = 5)	46.10%
Random forest model	77.57%
Unscaled MLP model	67.13%
Standard scaled MLP model	76.01%
Audio data stream model	62.88%
AF dual-stream CNN	98.03%

ResNet18 and VGGNet16 have been used previously for facial fatigue feature extraction [33]. Inspired by this, we selected related models for testing with the air traffic controller facial fatigue dataset. The experimental results are presented in Table 7.

Table 7. Comparison of facial detection model accuracies.

Model Name	Accuracy
VGG19	35.94%
VGG16	96.81%
ResNet50	97.82%
LeNet	95.01%
Facial data stream model	96.0%
AF dual-stream CNN	98.03%

The experimental results show that our dual-stream network exhibited high accuracy, 0.21% higher than that for ResNet50. In addition, our network converged markedly faster during training, requiring only 50 iterations to converge.

As indicated in Table 8, although the number of parameters of our model was almost the same as that for ResNet50, the number of iterations required for convergence was only 20% of those for ResNet50. This indicates that our network model converges rapidly, so it is particularly well suited to air traffic controller fatigue detection. Table 8 lists the values of four evaluation parameters for our AF dual-stream CNN model.

Table 8. Comparison of the numbers of detection model parameters and iterations required for convergence.

Model Name	Number of Nontrainable Parameters	Number of Trainable Parameters	Number of Iterations
VGG19	0	21,601,219	1000
VGG16	0	16,291,523	300
ResNet50	53,120	23,534,467	1200
LeNet	0	3,627,573	200
AF dual-stream CNN	0	23,582,979	50

We use four metrics, Precision, F1 Score, Recall, and Support, to evaluate the recognition results of our model, as shown in Table 9.

Table 9. Detection performance of our AF dual-stream CNN.

Label	Precision	F1 Score	Recall	Support
Severe fatigue	0.98	0.99	1.00	97
Mild fatigue	1.00	0.99	0.98	129
No fatigue	0.99	0.99	0.99	95

The comparative experimental results for the audio data show that the fatigue characteristics of audio were difficult to categorize into the three fatigue states. Even when using audio features that performed well in previous studies, the highest accuracy of the various test models for fatigue audio detection did not exceed 77.57%. The comparative experimental results for facial data show that our facial data stream model performed well but was not the best. In order to overcome the shortcomings of audio features and improve the detection accuracy of facial features, we combined facial data with audio data to judge the fatigue state of the air traffic controllers. The experimental results show that this combination approach resulted in the detection accuracy of our AF dual-stream CNN model increasing by 2.03%, reaching 98.03%.

4. Conclusions

This paper proposes an AF dual-stream CNN based on radio telephony data and facial data for the first time in the field of fatigue detection. The dual-stream convolutional network architecture designed in this study effectively utilizes the complementarity between multi-modal data sources. The experimental results show that the fatigue detection accuracy of our dual-stream network was 35.15% higher than that for using the radio telephony data alone and 2.03% higher than that for using the facial data alone, reaching 98.03%, which is better than the other algorithms and models tested in this study. In addition, during the training process, the neural networks assessed in experiments such as VGG16 also achieved detection accuracies as high as 96.81%, but the required number of iterations exceeded 300 times, and the training of ResNet50 required more than 1000 iterations; in contrast, our network model needs fewer than 50 training iterations to achieve convergence. The experimental results also show that networks such as VGG19, which have performed well in previous studies, did not perform well for our dataset, suggesting that such neural network models are not suitable for facial data, and their generalizability is not sufficient to support their implementation for classifying facial fatigue states. Our AF dual-stream CNN designed for fatigue detection effectively realizes the classification of controller fatigue states based on radio telephony data and facial data. The method in this paper can intervene in time when a controller shows fatigue, thereby contributing to the safe operation of flights.

The dual-stream convolutional neural network requires fewer iterations to reach convergence. In addition, we believe that there is still a better form of this structure, which can be further improved in the future. In our future work, we plan to focus on

solving two problems. Firstly, when an air traffic controller issues control instructions, their mouth movements will negatively impact the detection of their facial fatigue status, so how to further reduce the impact of such factors needs to be determined. Secondly, due to the diversity of audio features, it is necessary to identify those audio features and their combinations that are optimal for fatigue detection.

Author Contributions: Conceptualization, L.X. and Z.S.; methodology, S.M.; formal analysis, S.M. and Y.N.; validation, L.X. and S.M.; software Y.N.; resources, supervision, project administration, and funding acquisition, Z.S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the financial support from the National Natural Science Foundation of China (grant no. U2233208).

Informed Consent Statement: Informed Consent was obtained from all subjects involved in this study.

Data Availability Statement: The data of this study are available from the corresponding author upon request.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Piao, Q.; Xu, X.; Fu, W.; Zhang, J.; Jiang, W.; Gao, X.; Chen, Z. Fatigue Index of ATC in Number Recognition Task. In *Man-Machine-Environment System Engineering, Proceedings of the 21st International Conference on MMESE: Commemorative Conference for the 110th Anniversary of Xuesen Qian's Birth and the 40th Anniversary of Founding of Man-Machine-Environment System Engineering, Beijing, China, 23–25 October 2021*; Springer: Singapore, 2022; pp. 255–259.
2. Joseph, B.E.; Joseph, A.M.; Jacob, T.M. Vocal fatigue—Do young speech-language pathologists practice what they preach? *J. Voice* **2020**, *34*, 647.e1–647.e5. [CrossRef]
3. Kelly, D.; Efthymiou, M. An analysis of human factors in fifty controlled flight into terrain aviation accidents from 2007 to 2017. *J. Saf. Res.* **2019**, *69*, 155–165. [CrossRef] [PubMed]
4. Abd-Elfattah, H.M.; Abdelazeim, F.H.; Elshennawy, S. Physical and cognitive consequences of fatigue: A review. *J. Adv. Res.* **2015**, *6*, 351–358. [CrossRef] [PubMed]
5. Zhang, X.; Yuan, L.; Zhao, M.; Bai, P. Effect of fatigue and stress on air traffic control performance. In Proceedings of the 2019 5th International Conference on Transportation Information and Safety (ICTIS), Liverpool, UK, 14–17 July 2019; IEEE: Toulouse, France, 2019; pp. 977–983.
6. Devi, M.S.; Bajaj, P.R. Driver fatigue detection based on eye tracking. In Proceedings of the 2008 First International Conference on Emerging Trends in Engineering and Technology, Bursa, Turkey, 30 November–2 December 2017; IEEE: Toulouse, France, 2008; pp. 649–652.
7. Saradadevi, M.; Bajaj, P. Driver fatigue detection using mouth and yawning analysis. *Int. J. Comput. Sci. Netw. Secur.* **2008**, *8*, 183–188.
8. Azim, T.; Jaffar, M.A.; Mirza, A.M. Automatic fatigue detection of drivers through pupil detection and yawning analysis. In Proceedings of the 2009 Fourth International Conference on Innovative Computing, Information and Control (ICICIC), Kaohsiung, Taiwan, 7–9 December 2019; IEEE: Toulouse, France, 2009; pp. 441–445.
9. Moujahid, A.; Dornaika, F.; Arganda-Carreras, I.; Reta, J. Efficient and compact face descriptor for driver drowsiness detection. *Expert Syst. Appl.* **2021**, *168*, 114334. [CrossRef]
10. Khan, S.A.; Hussain, S.; Xiaoming, S.; Yang, S. An Effective Framework for Driver Fatigue Recognition Based on Intelligent Facial Expressions Analysis. *IEEE Access* **2018**, *6*, 67459–67468. [CrossRef]
11. Liang, H.; Liu, C.; Chen, K.; Kong, J.; Han, Q.; Zhao, T. Controller fatigue state detection based on ES-DFNN. *Aerospace* **2021**, *8*, 383. [CrossRef]
12. Deng, Y.; Sun, Y. A method to determine the fatigue of air traffic controller by action recognition. In Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 14–16 October 2020; IEEE: Toulouse, France, 2020; pp. 95–97.
13. Li, K.; Wang, S.; Du, C.; Huang, Y.; Feng, X.; Zhou, F. Accurate fatigue detection based on multiple facial morphological features. *J. Sens.* **2019**, *2019*, 7934516. [CrossRef]
14. Zhang, J.; Chen, Z.; Liu, W.; Ding, P.; Wu, Q. A field study of work type influence on air traffic controllers' fatigue based on data-driven PERCLOS detection. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11937. [CrossRef]
15. Abtahi, S.; Hariri, B.; Shirmohammadi, S. Driver drowsiness monitoring based on yawning detection. In Proceedings of the IEEE International Instrumentation and Measurement Technology Conference, Hangzhou, China, 10–12 May 2011; pp. 1–4.
16. Devi, M.S.; Bajaj, P.R. Fuzzy based driver fatigue detection. In Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10–13 October 2010; pp. 3139–3144.

17. Li, K.; Gong, Y.; Ren, Z. A fatigue driving detection algorithm based on facial multi-feature fusion. *IEEE Access* **2020**, *8*, 101244–101259. [CrossRef]
18. de Vasconcelos, C.A.; Vieira, M.N.; Kecklund, G.; Yehia, H.C. Speech analysis for fatigue and sleepiness detection of a pilot. *Aerosp. Med. Hum. Perform.* **2019**, *90*, 415–418. [CrossRef]
19. Wu, N.; Sun, J. Fatigue Detection of Air Traffic Controllers Based on Radiotelephony Communications and Self-Adaption Quantum Genetic Algorithm Optimization Ensemble Learning. *Appl. Sci.* **2022**, *12*, 10252. [CrossRef]
20. Shen, Z.; Pan, G.; Yan, Y. A high-precision fatigue detecting method for air traffic controllers based on revised fractal dimension feature. *Math. Probl. Eng.* **2020**, *2020*, 4563962. [CrossRef]
21. Sun, H.; Jia, Q.; Liu, C. Study on voice feature change of radiotelephony communication under fatigue state. *China Saf. Sci. J.* **2020**, *30*, 158.
22. Shen, Z.; Wei, Y. A high-precision feature extraction network of fatigue speech from air traffic controller radiotelephony based on improved deep learning. *ICT Express* **2021**, *7*, 403–413. [CrossRef]
23. Dobrišek, S.; Gajšek, R.; Mihelič, F.; Pavešić, N.; Štruc, V. Towards efficient multi-modal emotion recognition. *Int. J. Adv. Robot. Syst.* **2013**, *10*, 53. [CrossRef]
24. Panda, R.; Malheiro, R.M.; Paiva, R.P. Audio features for music emotion recognition: A survey. *IEEE Trans. Affect. Comput.* **2020**, *14*, 68–88. [CrossRef]
25. Zaw, T.H.; War, N. The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017; IEEE: Toulouse, France, 2017; pp. 1–5.
26. Yuan, S.; Wang, Z.; Isik, U.; Giri, R.; Valin, J.-M.; Goodwin, M.M.; Krishnaswamy, A. Improved singing voice separation with chromagram-based pitch-aware remixing. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Toulouse, France, 2022; pp. 111–115.
27. Muda, L.; Begam, M.; Elamvazuthi, I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv* **2010**, arXiv:1003.4083.
28. Madsen, P.T. Marine mammals and noise: Problems with root mean square sound pressure levels for transients. *J. Acoust. Soc. Am.* **2005**, *117*, 3952–3957. [CrossRef]
29. Soundarya, M.; Karthikeyan, P.R.; Ganapathy, K.; Thangarasu, G. Automatic Speech Recognition using the Melspectrogram-based method for English Phonemes. In Proceedings of the 2022 International Conference on Computer, Power and Communications (ICCPC), Chennai, India, 14–16 December 2022; IEEE: Toulouse, France, 2022; pp. 270–273.
30. Hu, Y.; Liu, Z.; Hou, A.; Wu, C.; Wei, W.; Wang, Y.; Liu, M. On fatigue detection for air traffic controllers based on fuzzy fusion of multiple features. *Comput. Math. Methods Med.* **2022**, *2022*, 4911005. [CrossRef]
31. Yang, L.; Li, L.; Liu, Q.; Ma, Y.; Liao, J. Influence of physiological, psychological and environmental factors on passenger ship seafarer fatigue in real navigation environment. *Saf. Sci.* **2023**, *168*, 106293. [CrossRef]
32. Sun, J.; Sun, R.; Li, J.; Wang, P.; Zhang, N. Flight crew fatigue risk assessment for international flights under the COVID-19 outbreak response exemption policy. *BMC Public Health* **2022**, *22*, 1843. [CrossRef] [PubMed]
33. Jamshidi, S.; Azmi, R.; Sharghi, M.; Soryani, M. Hierarchical deep neural networks to detect driver drowsiness. *Multimed. Tools Appl.* **2021**, *80*, 16045–16058. [CrossRef]