frontiers | Frontiers in Neuroscience

TYPE Original Research
PUBLISHED 05 August 2022
DOI 10.3389/fnins.2022.947168

Check for updates

OPEN ACCESS

EDITED BY
Lin Shu,
South China University of Technology,
China

REVIEWED BY
Hasan AL-Nashash,
American University of Sharjah, United Arab Emirates
Giuseppe D'Avenio,
National Institute of Health (ISS), Italy

*CORRESPONDENCE
Ling He
ling.he@scu.edu.cn

SPECIALTY SECTION
This article was submitted to Perception Science, a section of the journal Frontiers in Neuroscience

RECEIVED 18 May 2022
ACCEPTED 11 July 2022
PUBLISHED 05 August 2022

CITATION
Zhang J, Yin H, Zhang J, Yang G, Qin J and He L (2022) Real-time mental stress detection using multimodality expressions with a deep learning framework.
Front. Neurosci. 16:947168.
doi: 10.3389/fnins.2022.947168

Real-time mental stress detection using multimodality expressions with a deep learning framework

Jing Zhang[1], Hang Yin[1], Jiayu Zhang[1], Gang Yang[1], Jing Qin[2] and Ling He[1]*

[1]College of Biomedical Engineering, Sichuan University, Chengdu, China, [2]Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China

Mental stress is becoming increasingly widespread and gradually severe in modern society, threatening people's physical and mental health. To avoid the adverse effects of stress on people, it is imperative to detect stress in time. Many studies have demonstrated the effectiveness of using objective indicators to detect stress. Over the past few years, a growing number of researchers have been trying to use deep learning technology to detect stress. However, these works usually use single-modality for stress detection and rarely combine stress-related information from multimodality. In this paper, a real-time deep learning framework is proposed to fuse ECG, voice, and facial expressions for acute stress detection. The framework extracts the stress-related information of the corresponding input through ResNet50 and I3D with the temporal attention module (TAM), where TAM can highlight the distinguishing temporal representation for facial expressions about stress. The matrix eigenvector-based approach is then used to fuse the multimodality information about stress. To validate the effectiveness of the framework, a well-established psychological experiment, the Montreal imaging stress task (MIST), was applied in this work. We collected multimodality data from 20 participants during MIST. The results demonstrate that the framework can combine stress-related information from multimodality to achieve 85.1% accuracy in distinguishing acute stress. It can serve as a tool for computer-aided stress detection.

KEYWORDS

stress detection, objective indicators, multimodality fusion, deep learning, matrix eigenvector

Introduction

Stress is an individual's adaptation response to internal or external threats (Freeman, 1986; Mitra, 2008). It can affect people's daily performance, memory, and decision-making abilities (Sharma and Gedeon, 2012; Nigam et al., 2021). Acute stress occurs when people are faced with urgent tasks such as mental arithmetic, academic exams, or public speaking (Allen et al., 2017). It usually disappears when the urgent task is over. If acute stress continues to permeate a person's life, it can result in decreased physical and mental health and even lead to immune system disorders, cardiovascular disease, depression, or other diseases (Sauter et al., 1990; Segerstrom and Miller, 2004; van Praag, 2004; Heraclides et al., 2012; Beanland et al., 2013; Steptoe and Kivimäki, 2013). In modern society, stress has become increasingly widespread and severe. The European Union has established it as one of the most common causes of health problems (Wiegel et al., 2015).

To avoid harm to people caused by chronic or acute stress, it is essential to detect people's stress state as early as possible and prevent the adverse effects of stress on people. Psychological evaluation of stress can be used to detect an individual's stress state (Alberdi et al., 2016). Stress is assessed by filling out a questionnaire or talking to a psychologist. Since psychological evaluation is instantaneous and subjective, it often leads to false or even incorrect stress detection and is unable to meet the requirements of real-time detection (Ren et al., 2019). In contrast, using objective indicators such as physiological signals and behavioral information to detect stress is not affected by subjective influence (Setz et al., 2010; Cinaz et al., 2011; McDuff et al., 2012; Wei, 2013; Sharma et al., 2021).

When people are under stress, the autonomic nervous system (ANS) is stimulated and regulates involuntary body functions (Tsigos and Chrousos, 2002). As a result of changes in involuntary body functions, the electrocardiogram (ECG), voice, and facial expressions of people are affected. ECG is the physiological signal that can record cardiac activity. As regulated by the ANS, during stress, the heart rate increases, and the heartbeat's standard deviation becomes larger (De Rosa, 2004; de Santos Sierra et al., 2011). These changes can be presented by ECG. Dominated by the ANS during stress, the pitch and speaking rate of voice are affected, while the energy and spectral characteristics of voice also change. The mean value, standard deviation, range of pitch increase, and jitter of pitch decrease when people are under stress. The spectral centroid goes up, and energy is concentrated in higher frequency bands (Lu et al., 2012). Likewise, affected by stress, facial expressions involving the eyes, mouth, and cheeks are different from calm (Liao et al., 2005; Sharma and Gedeon, 2012; Sundelin et al., 2013; Pampouchidou et al., 2016). The overall changes in these multiple facial regions constitute changes in facial expressions.

In recent years, multimodality fusion methods have received increasing attention. It has been widely used in computer-aided diagnosis and performs better prediction than single-modality-based methods (Yang et al., 2013; Vidya et al., 2015; Zhu et al., 2020). Since the ECG, voice, and facial expressions describe stress changes in a different way and are jointly affected by the ANS (Giannakakis et al., 2017). Fusing these multimodalities can detect the stress state from multiple aspects.

Deep learning technology has shown excellent performance in many fields (LeCun et al., 2015; Hatcher and Yu, 2018). Different from the handcrafted feature engineering methods, it automatically extracted the features of input through the deep learning network to minimize the feature extraction process and achieve better generalization ability. Due to the advantages of deep learning, a growing number of researchers are trying to use deep learning technology to detect stress (Jin et al., 2016; Hwang et al., 2018; Winata et al., 2018). Convolutional neural networks (CNNs) are an attractive way to distinguish different classes of inputs in deep learning technology. CNNs can extract features in multiple dimensions of the input, among which 2D-CNN can capture the global and local spatial information, and 3D-CNN can also capture the temporal information. Studies have proven that CNNs are effective for stress detection (Jin et al., 2016; Hwang et al., 2018; Winata et al., 2018), but the potential of using CNNs that fuse multimodality for stress detection remains to be explored.

In this work, a real-time deep learning framework that fused ECG, voice, and facial expressions for acute stress detection is proposed. Furthermore, we designed the temporal attention module (TAM) to find the keyframes related to stress detection in facial expressions. The proposed framework avoids complicated feature extraction and only requires simple preprocessing. The contributions of our work can be summarized as follows:

(1) This work proposes a deep learning framework that combines ECG, voice, and facial expressions for acute stress detection. The fusion method is based on the matrix eigenvector, which achieves 85.1% detection accuracy.

(2) The proposed framework utilizes TAM. The TAM assigns different learnable weights to different frames of facial expressions to highlight the distinguishing temporal representation for facial expressions about stress.

This research is organized as follows. The section "Materials and methods" introduces multimodality data acquisition, preprocessing, and the real-time deep learning framework. Section "Results" shows the results of our experiment, and Sections "Discussion" and "Conclusion" present the discussion and conclusion of our research, respectively.

Materials and methods

Materials

To collect multimodality data from people under stress, it is necessary to stimulate stress in participants by designed experiments (Setz et al., 2010; Tomova et al., 2016; Smets et al., 2018; Stepanovic et al., 2019). Many different stress-induced methods have been validated to stimulate stress. The most commonly used experimental paradigms are the Stroop Color-Word Interference Test and the Montreal Imaging Stress Task (MIST; Aitken, 1969; Lovallo, 1975; Kirschbaum et al., 1993; Renaud and Blondin, 1997; Dedovic et al., 2005; Reinhardt et al., 2012; Smeets et al., 2012; Tanosoto et al., 2012).

MIST is the gold standard experiment for stimulating stress. As a well-established psychological experiment employed in stress assessment, it has been proven to put people into a stress state by measuring the amount of cortisol in their saliva (Lederbogen et al., 2011; Kiem et al., 2013; Sioni and Chittaro, 2015). To date, a large number of studies on stress have been carried out on the basis of MIST and its modified experiments (Boehringer et al., 2015; Chung et al., 2016; Wheelock et al., 2016; Gossett et al., 2018; Hakimi and Setarehdan, 2018; Li et al., 2018; Xia et al., 2018; Noack et al., 2019; Perez-Valero et al., 2021). The MIST is a computer-based standardized psychological experimental designed to assess the effects of psychological stress on people's physiology and behavior (Dedovic et al., 2005). To obtain multimodality data from participants in a stressful state, this work used MIST to induce people to be under stress.

Participants

Twenty right-handed participants (11 males, 9 females, mean age = 22.75, SEM = 0.13, age-range 20–25 years, 20 Chinese) participated in the MIST to stimulate psychological stress. All the participants were participating in MIST for the first time. The overall flow of MIST was introduced before the experiment started.

Data collection

MIST is a computer-based psychological experimental paradigm that mainly includes (1) the calm stage, (2) the control stage, (3) the experimental stage, and (4) the recovery stage.

In the calm stage, the participants read the equation with the answer. In the control stage, the participant clicks on the correct answer in the program and reads out the calculation and the result, and the screen will display correct or incorrect. During the experimental stage, the participants perform calculations with time constraints, and the program adaptively adjusts the time constraints and difficulty. If the participants correctly solve three arithmetic tasks in a row, the program will reduce the time constraints and raise calculation difficulty. Both the control stage and the experimental stage can cause psychological stress in the participants. The recovery helps participants return to a calm state. After the MIST experiment, each participant was asked to fill out a stress questionnaire. **Figure 1** shows the MIST process.

This work used the MIST program written and deployed using JDK 8u66 for Windows. The program can automatically create arithmetic tasks including addition, subtraction, multiplication, and division. There are five categories of difficulty for calculation problems, the two easiest of which are the addition or subtraction of 2 or 3 one-digit integers. The two classes of medium difficulty contain 3 or 4 integers and allow for multiplication. For the most difficult category, the calculation includes addition, subtraction, multiplication, and division of four integers. The answers to all computational tasks are integers between 0 and 9.

The multimodality data collection platform used in this work includes two computers (one for the MIST experiment and one for collecting data), a physiological signal acquisition device Biopac MP160, and a Sony video camera FDR-AX700. The MP160 contains the participant's ECG signal through a wireless transmission module and transmits data to the collecting computer through the network cable. The ECG signal is acquired by three-electrode leads, two electrodes placed symmetrically in the fourth or fifth rib region, and one electrode placed in the right upper chest area, where the sampling frequency is 2000 HZ. The camera captures the facial expressions and voice of the participants during the MIST experiment and sends them to the collecting computer, where the resolution of the video is 1920*1080 and 30 fps. The camera and sensor are turned on at the same time, aligning the data according to the data and video length. The platform diagram is shown in **Figure 2**.

The process of participants completing one equation is considered as a piece of data. A piece of the sample contains participants' facial expressions, ECG, and voice. The samples were labeled according to the stage of MIST. In this way, 1271 samples were acquired, including 531 labeled "calm" and 740 labeled "stress."

Methods

The real-time deep learning framework proposed in this work used ECG, voice, and facial expressions for stress detection. Each modal was preprocessed before being input into the framework. ECG and voice were converted into the form that represents their time-frequency changes, and facial expressions were extracted from the collected video. Then the multidimensional features of each modality were extracted through the deep learning framework. The fully connected layers in the framework obtained the information about the stress state, and the framework fused them into a global matrix for stress detection. An overview of multimodality stress detection in this work is shown in **Figure 3**.

Data preprocessing

Facial expressions preprocessing

This work removed background information to isolate the facial area, which can avoid being disturbed by irrelevant

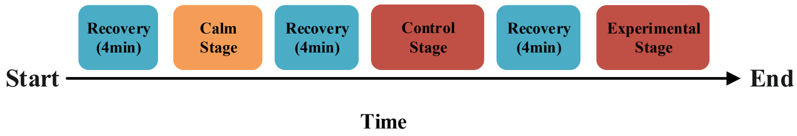

FIGURE 1
The Montreal imaging stress task process.

information of surrounding noise and clutter in different scenarios of reality. Video $\mathbf{V_i}$ was framed into sequence of image frames $\mathbf{V_i} = (\mathbf{frame_{i_1}}, \mathbf{frame_{i_1}}, ..., \mathbf{frame_{i_n}})$, and the face area was detected by MTCNN [45] on each image frame and then aligned. $\mathbf{Face_i} = (\mathbf{face_{i_1}}, \mathbf{face_{i_1}}, ..., \mathbf{face_{i_n}})$ denotes a sequence of face images detected from sequence $\mathbf{V_i}$.

Electrocardiogram preprocessing

The original ECG signal contains high-frequency and electrical noise. And intercepting the ECG signal according to temporal leads to the start position and the end position of the heartbeat in different samples being inconsistent. We preprocessed the original ECG signal below to solve these problems. The preprocess of ECG is shown in **Figure 4**.

(1) Denoising: The ECG signal was passed through a notch filter with a cutoff frequency of 50 Hz to eliminate the interference caused by the industrial frequency current. The energy of the ECG signal is concentrated in the frequency band of less than 50Hz. Filtering out the ECG signal higher than 50Hz will not affect the expression of ECG changes. Then a pass filter with a cutoff frequency of (0.5, 50) Hz was used to reduce the influence of electrode noise, muscle noise, and baseline wander noise on the ECG signal. After that, the ECG signal was standardized to eliminate amplitude scaling in the heartbeat cycles.

(2) Heartbeat relocating: Due to the interception according to the temporal causes, the start position and the end position of the heartbeat in different samples were inconsistent. In this work, the start position and end position of the ECG signal were relocated based on Niu's work (Niu et al., 2020). It takes the middle temporal position between the first **R** wave in the sample and its preceding **R** wave in the collected ECG data as the start position. The middle temporal position between the last **R** wave in the sample and the **R** wave after it in the collected ECG data is considered the end position. The short-term autocorrelation function is used to calculate the temporal distance between two **R** waves (Piotrowski and Różanowski, 2010).

First, We detected the temporal position of each **R** wave in the ECG signal through a sliding window with a threshold. The temporal position of the **R** waves was recorded as$\mathbf{R_1}, ..., \mathbf{R_N}$. Second, we reversed the ECG signal 2 s before the $\mathbf{R_1}$ and calculate the short-term autocorrelation function as $\mathbf{X_1(n)}$. The short-term autocorrelation function of the ECG signal 2 s after the $\mathbf{R_5}$ was calculated as $\mathbf{X_N(n)}$. $\mathbf{X_1(n)}$ and $\mathbf{X_N(n)}$ were clipped using thresholds $\boldsymbol{\alpha} = \mathbf{0.1}$ and $\boldsymbol{\beta} = \mathbf{-0.1}$ with the following formula:

$$\mathbf{X(n)} = \begin{cases} \mathbf{x(n) - \alpha, x(n) > \alpha} \\ \mathbf{0, \beta \leq x(n) \leq \alpha} \\ \mathbf{x(n) - \beta, x(n) < \beta} \end{cases}$$

The temporal between the first sample and the maximum sample of the first harmonic in $\mathbf{X_1(n)}$ represents the temporal distance $\mathbf{T_1}$ between $\mathbf{R_1}$ and its preceding **R** wave. We take the middle temporal position of $\mathbf{T_1}$ as the start position. The temporal between the first sample and the maximum sample of the first harmonic in $\mathbf{X_N(n)}$ represents the temporal distance $\mathbf{T_2}$ between $\mathbf{R_n}$ and the **R** wave after it. The middle temporal position of T2 was taken as the end position of the heartbeat.

(3) Visualization: The relocated heartbeat were converted into the form of an image. The vertical axis represents the amplitude of the heartbeat and the horizontal axis represents the temporal of the heartbeat.

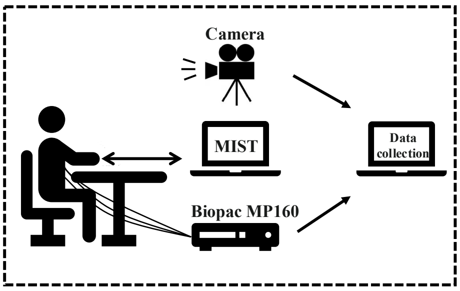

FIGURE 2
Data collection platform.

Voice preprocessing

The Mel spectrogram calculated in the time-frequency domain analysis contains time and frequency information of voice. It converts the linear frequency scale into a logarithmic scale and represents the distribution of signal energy on the Mel-scale frequency, which is similar to human hearing. Mel spectrogram can intuitively show the spectral changes of voice over time. Therefore, we convert the voice into the Mel spectrogram.

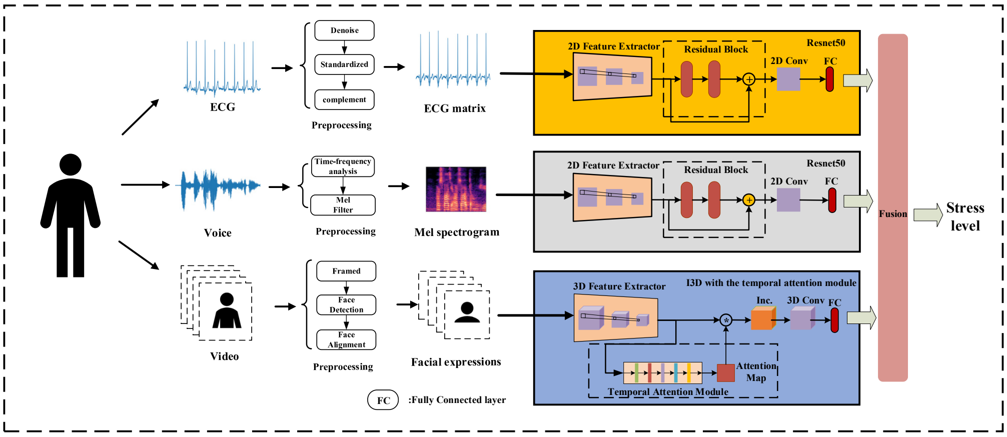

FIGURE 3
Multimodality stress detection.

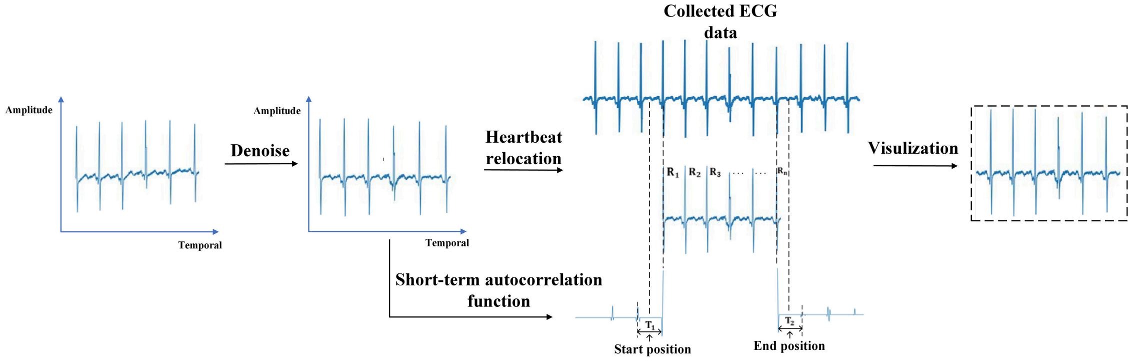

FIGURE 4
Electrocardiogram preprocessing.

After the voice data are divided into each sample, we pre-emphasis the data. Then, the Hamming window of 30ms length is used to frame the data with 15 ms overlap.

After that, we calculate energy density using the short-time Fourier transform (STFT) and transform the frequency into Mel-scale band to extract Mel spectrogram.

Real-time deep learning framework for stress detection

The real-time deep learning framework was developed by combining ResNet50 (He et al., 2016) and I3D with the temporal attention module. ResNet50 extracts the global and local features of the ECG matrix and Mel spectrogram through identity mapping. I3D with the temporal attention module learns the spatiotemporal changes in facial expressions, and the temporal attention module enables I3D to extract important temporal features.

The stress state information in their fully connected layers was combined into a global matrix, leading to a multimodal stress information representation. The multimodality information about stress was fused for stress detection based on matrix eigenvectors.

ResNet50

Converting the ECG and voice signals into three-dimensional matrices can represent higher-order and nonlinear characteristics of the signals. This work used Resnet in 2D-CNN to extract multiple features from the ECG and Mel spectrogram. ResNet avoids the problem of gradient disappearance and explosion in traditional 2D-CNN through shortcut connections. The 2D convolution strategy of ECG and Mel spectrogram input in this work is shown in Figure 5.

The 2D-CNN has a convolution layer composed of 2D convolution kernels. The convolutional layer can extract features in multiple dimensions to obtain the feature representation of the internal structure of the ECG and Mel spectrogram by scanning them with the 2D convolution

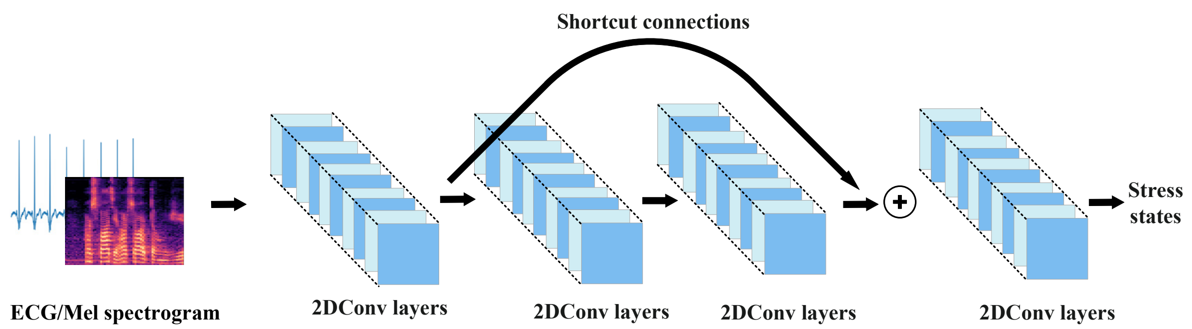

FIGURE 5
2D convolution strategy for electrocardiogram and Mel spectrogram input in this work.

kernels and reducing the number of parameters through local connectivity and parameter sharing. The 2D convolution can be expressed as:

$$\mathbf{v_{ij}^{xy}} = \mathbf{f(b_{ij}} + \sum_{\mathbf{m}} \sum_{\mathbf{p=0}}^{\mathbf{P_i-1}} \sum_{\mathbf{q=0}}^{\mathbf{Q_i-1}} \mathbf{W_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)}})$$

where $\mathbf{v_{ij}^{xy}}$ is the **i** convolution result at the **j** position in feature map $(\mathbf{x}, \mathbf{y})$ of the layer; **f** is the activation function Rectified Linear Unit (ReLU; Nair, 2010); $\mathbf{b_{ij}}$ is the deviation of the feature map; **m** is the index of the feature map in layer $\mathbf{i-1}$; $\mathbf{Q_i}$, $\mathbf{P_i}$ is the height and width of the convolution kernel; and $\mathbf{W_{ijm}^{pq}}$ represents the value at the position of the feature map.

In traditional 2D-CNN, when the network structure becomes very deep, there will be unavoidable problems of gradient disappearance or explosion, and the problem of accuracy saturation or decline. This causes such networks to be unable to capture the overall stress information in the input. Resnet avoids the problems caused by a network structure that is too deep through the residual block with shortcut connections. The residual block connects the inputs in the lower layers and high layers which converts the input maps into identity maps.

This work used ResNet50 to extract effective representations of stress in ECG and Mel spectrogram input. ResNet50 has half the floating-point operations (FLOPs) of ResNet101, and only 5% more than ResNet34. It reduces the number of FLOPs while satisfying the accuracy requirements. The overall architecture of ResNet50 is shown in **Figure 6**.

The residual block in **Figure 6** is defined as:

$$\mathbf{y} = \mathbf{F}\left(\mathbf{x}, \{\mathbf{W_i}\}\right) + \mathbf{X}$$

where **X** is the input of weight layer; **ReLU** is the activation function; $\mathbf{F(X, \{W_i\})}$ is the output after three convolution layers. Identity mapping adds $\mathbf{F(X)}$ and **X** as the input **y** to the next residual block.

I3D With the Temporal Attention Module

As mentioned above, detecting stress by facial expressions requires comparing temporal changes in multiple facial regions, including the eyes, mouth, and cheeks. This work used the inflated 3D-CNN (I3D) with the temporal attention module to learn the overall changes in facial expressions during stress. The I3D extracts the features of stress in facial expressions and the temporal attention module tells I3D which frames are important. The 3D convolution strategy for facial expressions input in this work is shown in **Figure 7**.

The 3D-CNN has a 3D convolution kernel that can analyze successive frames of facial expressions and capture spatiotemporal features of facial expressions. The 3D convolution can be expressed as:

$$\mathbf{v_{km}^{xyz}} = \mathbf{f(b_{km}} + \sum_{\mathbf{p=0}}^{\mathbf{P_k-1}} \sum_{\mathbf{q=0}}^{\mathbf{Q_k-1}} \sum_{\mathbf{r=0}}^{\mathbf{R_k-1}} \mathbf{w_{kmn}^{pqr} u_{(k-1)n}^{(x+p)(y+q)(z+r)}})$$

where $\mathbf{v_{km}^{xyz}}$ is the position **k** in the **m** feature map of the $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ layer; **f** is the loss function; **u** is the input from the $\mathbf{k-1}$ to **k** layer; $\mathbf{P_k}$, $\mathbf{Q_k}$, $\mathbf{R_k}$ are the width height and depth of the convolution kernel size; $\mathbf{b_{km}}$ is the deviation of the $\mathbf{w_{kmn}^{pqr}}$ feature map.

In traditional 3D-CNN (C3D), each layer generally uses a single-size convolution kernel and forward propagation, which cannot extract overall features. This results in the entire facial expression being ineffectively represented. The inflated 3D-CNN (I3D; Carreira and Zisserman, 2017) combines the advantages of GoogLenet (Szegedy et al., 2015) and 3D-CNN. It uses the inflated inception module for feature extraction. The inflated inception module uses convolution kernels of different sizes to extract features, and finally concatenates them to increase the network's ability to extract overall features. Therefore, for facial expressions, the corresponding features of stress in multiple facial regions can be extracted by I3D.

During facial expression changes, subtle changes tend to last only a few frames, so not all frames are equally important

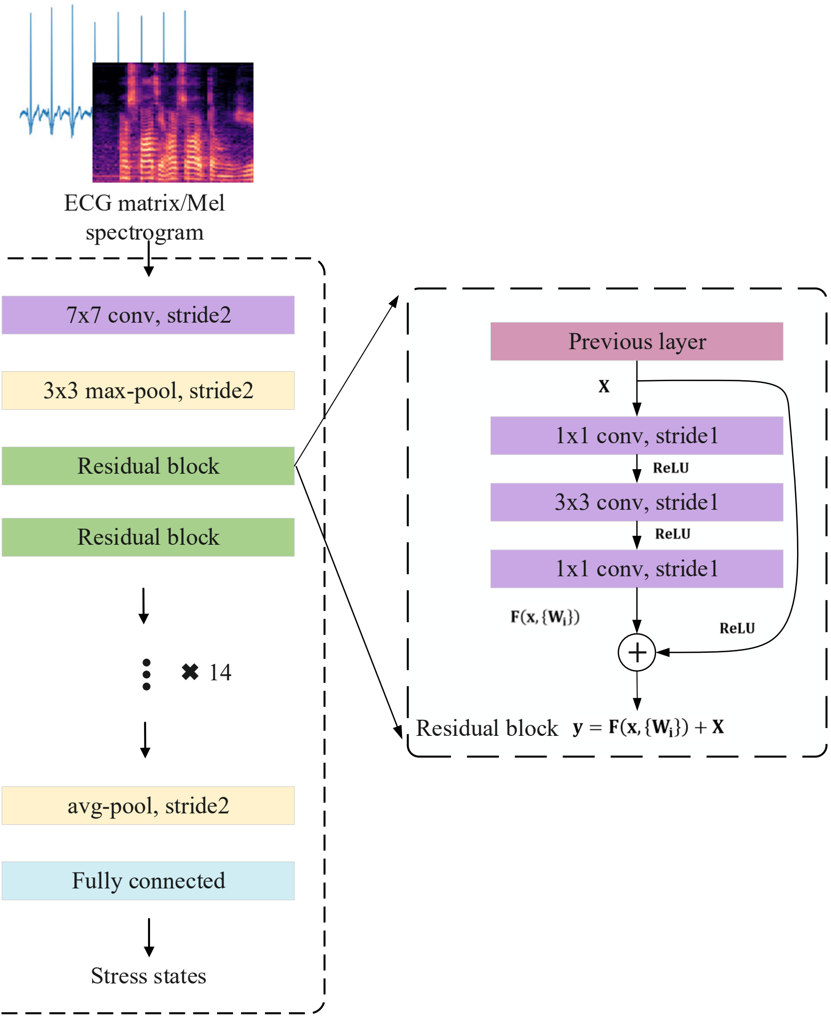

FIGURE 6
The overall architecture of the ResNet50.

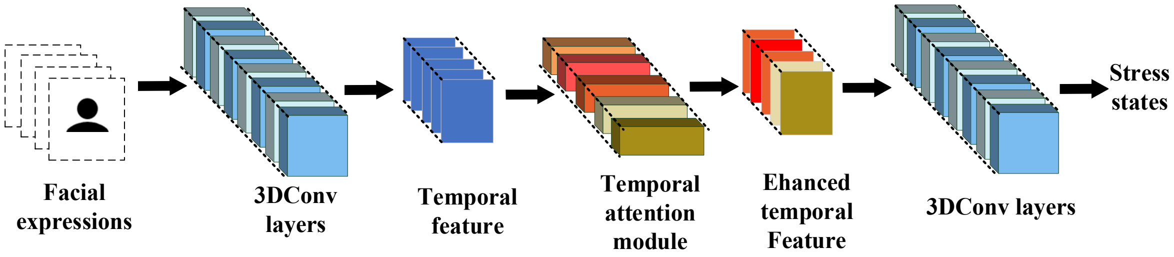

FIGURE 7
3D convolution strategy for facial expressions input in this work.

for distinguishing facial expressions. The identical scale is used to pool temporal information in traditional I3D, which makes important frames lost and trapped in local details.

To enhance the global perception of temporal information in I3D and avoid getting caught in local temporal details, this work proposed the temporal attention module (TAM) for the I3D layers. In TAM, global pooling is calculated for each input frame. Then two fully connected layers and a sigmoid function are used to generate a temporal attention map, which is finally combined with the multiplication of the original feature maps to change the proportion of temporal information captured by the initial layer of I3D. It makes I3D highlight the distinguishing features while ignoring interfering features when extracting the temporal information of facial expressions. The overall architecture of I3D with TAM is shown in Figure 8.

Formally, a statistic $\mathbf{z} \in \mathbf{R}^{\mathbf{T}}$ is generated by shrinking $\mathbf{u} \in \mathbf{R}^{\mathbf{C} \times \mathbf{H} \times \mathbf{W}}$ $(\mathbf{F_s(.)})$ through its channel and spatial information $\mathbf{C} \times \mathbf{H} \times \mathbf{W}$:

$$\mathbf{z_t} = \mathbf{F_s}(\mathbf{u_t}) = \frac{1}{\mathbf{C} \times \mathbf{H} \times \mathbf{W}} \sum_{i=1}^{C} \sum_{j=1}^{H} \sum_{k=1}^{W} \mathbf{u_t}(\mathbf{i}, \mathbf{j}, \mathbf{k})$$

To use the shrinking information, we follow it with a second operation $\mathbf{F_{ex}(.)}$. We want to ensure that different frames are allowed to be emphasized, so we choose sigmoid activation as a simple gating mechanism:

$$\mathbf{S} = \mathbf{F_{ex}}(\mathbf{z}, \mathbf{W}) = \sigma\left(\mathbf{g}(\mathbf{z}, \mathbf{W})\right) = \sigma(\mathbf{W_2}\delta(\mathbf{W_1 z}))$$

where δ refers to the ReLU function. $\mathbf{W_1} \in \mathbf{R}^{\frac{\mathbf{T}}{\mathbf{r}} \times \mathbf{T}}$ and $\mathbf{W_2} \in \mathbf{R}^{\mathbf{T} \times \frac{\mathbf{T}}{\mathbf{r}}}$, where $\mathbf{r}$ is the reduction ratio. The final output of the TAM is acquired by multiplying $\mathbf{S}$ and $\mathbf{u}$:

$$\mathbf{Output} = \mathbf{F_{scale}}(\mathbf{u_t}, \mathbf{S_t}) = \mathbf{S_t u_t}$$

Multimodality fusion

This work proposed a fusion method based on the eigenvector corresponding to the largest eigenvalue of the matrix. In our method, the posterior probability information of each stage for each modality is composed as a vector. The vectors of the three modalities are composed into the stress information matrix. Our method uses the normalized eigenvector corresponding to the largest eigenvalue of the matrix as the weight to fuse the multimodality information.

After Resnet50 and I3D with TAM extracted input features about stress, their fully connected layer can obtain the posterior probability information of the stress stage of the input. For each modality, the probability of each stage $\mathbf{P_{calm}}, \mathbf{P_{control}}, \mathbf{P_{experimental}}$ derived from this modality can be composed as a vector, and the vectors of the three modalities $\mathbf{V_{voice}}, \mathbf{V_{face}}, \mathbf{V_{ECG}}$ can be composed as a matrix, which contains probabilistic information about each stage of each modal. We define this matrix as the stress information matrix $\mathbf{M_{global}}$.

$$\mathbf{M_{global}} = (\mathbf{V_{Voice}}, \mathbf{V_{Face}}, \mathbf{V_{ECG}}) = \begin{pmatrix} \mathbf{P_{calm\_V}}, & \mathbf{P_{calm\_F,}} & \mathbf{P_{calm\_E}} \\ \mathbf{P_{control\_V}}, & \mathbf{P_{control\_F}}, & \mathbf{P_{control\_E}} \\ \mathbf{P_{experimental\_V}}, & \mathbf{P_{experimetal\_F}}, & \mathbf{P_{experimental\_E}} \end{pmatrix}$$

In $\mathbf{M_{global}}$, the eigenvalues of $\mathbf{M_{global}}$ indicate how much the probability is scaled, and the eigenvectors of $\mathbf{M_{global}}$ indicate the direction of the probability. Compared with other eigenvalues and their corresponding eigenvectors, the largest eigenvalue and its corresponding eigenvector indicate that the probability in this direction is amplified to the greatest extent, that is, the probability of this matrix in this direction is the largest. Given this property of them, we use the eigenvector corresponding to the largest eigenvalue in $\mathbf{M_{global}}$ to calculate the weight vector to fuse the multimodality information from $\mathbf{M_{global}}$ for stress detection.

In $\mathbf{M_{global}}$, the probability of the three stages of MIST is included. The eigenvector $\mathbf{W_{max}}$ represents the eigenvector corresponding to the largest eigenvalue in $\mathbf{M_{global}}$. We normalize $\mathbf{W_{max}}$ as the weight vector $\mathbf{W_{weight}}$. $\mathbf{w_{calm}}, \mathbf{w_{control}}, \mathbf{w_{experimental}}$ represent the weight values of the calm, control, and experimental stages, respectively.

$$\mathbf{M_{global}} \rightarrow \mathbf{W_{max}} \xrightarrow{\text{normalize}} \mathbf{W_{weight}} = (\mathbf{w_{calm}}, \mathbf{w_{control}}, \mathbf{w_{experimental}})$$

After obtaining the weight values, the weight values are used to construct a diagonal weight matrix $\mathbf{W}$.

$$\mathbf{W} = \begin{pmatrix} \mathbf{w_{calm}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w_{control}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w_{experimental}} \end{pmatrix}$$

The cross-modal global matrix $\mathbf{M_{global}}$ is multiplied with the weight matrix to obtain the weighted matrix $\mathbf{W_{global}}$.

$$\mathbf{W_{global}} = \mathbf{W} * \mathbf{M_{global}} = (\mathbf{w_{calm}}, \mathbf{w_{control}}, \mathbf{w_{experimental}})$$

$$(\mathbf{w_{calm}}, \mathbf{w_{control}}, \mathbf{w_{experimental}}) = \begin{pmatrix} \mathbf{w_{calm\_V}}, & \mathbf{w_{calm\_F,}} & \mathbf{w_{calm\_E}} \\ \mathbf{w_{control\_V}}, & \mathbf{w_{control\_F}}, & \mathbf{w_{control\_E}} \\ \mathbf{w_{experimental\_V}}, & \mathbf{w_{experimental\_F}}, & \mathbf{w_{experimental\_E}} \end{pmatrix}$$

In $\mathbf{W_{global}}$, $\mathbf{w_{calm}} = (\mathbf{w_{calm\_V}}, \mathbf{w_{calm\_F}}, \mathbf{w_{calm\_E}})$ represents the weighted posterior probability of the calm stage. $\mathbf{w_{control}} = (\mathbf{w_{control\_V}}, \mathbf{w_{control\_F}}, \mathbf{w_{control\_E}})$ represents the weighted posterior probability of the control stage. $\mathbf{w_{experimental}} = (\mathbf{w_{experimental\_V}}, \mathbf{w_{experimental\_F}}, \mathbf{w_{experimental\_E}})$ represents the weighted posterior probability of the experimental stage. Whether the value of $\mathbf{W_{global}}$ $(\mathbf{w_i})$ is positive or negative, it's in the direction of the represented stage axis, the absolute size of its value represents the probability in this stage. Our method's output is the stress state corresponding to the absolute maximum in $\mathbf{W_{global}}$. The fusion process is shown in Figure 9.

$$\mathbf{Output} = \mathbf{max}\left[\left| \sum\nolimits_{i=1}^{3} \mathbf{w_{calmi\_i}}, \sum\nolimits_{i=1}^{3} \mathbf{w_{controli\_i}}, \sum\nolimits_{i=1}^{3} \mathbf{w_{experimental\_i}} \right|\right]$$

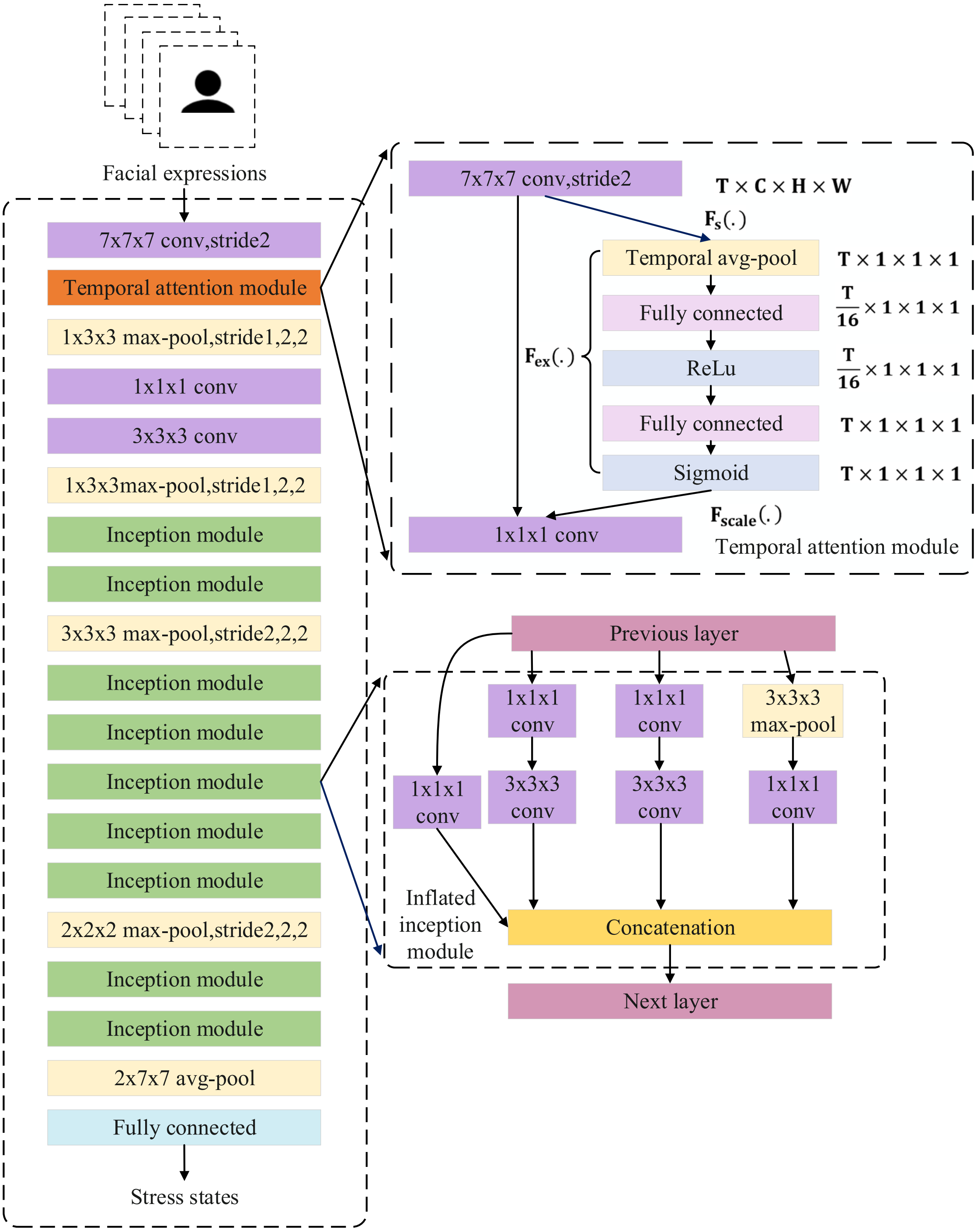

FIGURE 8
The overall architecture of the I3D with temporal attention module.

Results

The performance of the proposed deep learning framework for stress detection was evaluated with the collected dataset through four sets of experiments, including multimodality stress detection, single-modality-based stress detection, the effectiveness of the temporal attention module in I3D, and comparison with different models. The 10-fold-cross-validation method was utilized on the dataset for cross-validation to ensure the generalization ability of our method.

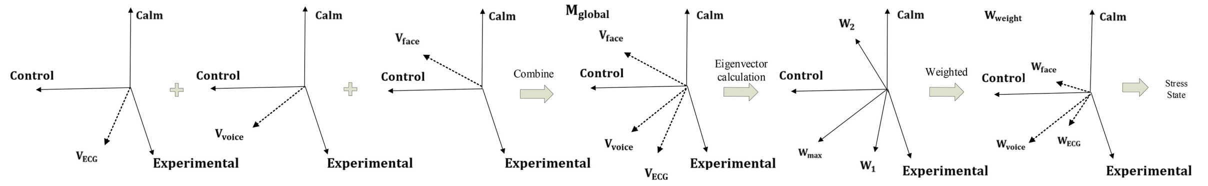

FIGURE 9
The fusion method based on matrix eigenvector.

Implementation details

This work randomly divided 80% of the data after leaving one-fold out into the training set and 20% into the validation set. The experimental environment is 64-bit Windows 10, GeForce RTX2070, and 16 GB memory. Our implementation is based on Python version 3.6.13 and PyTorch version 1.9.0 with CUDA version 11.4.

ResNet50 parameter settings

During training, the trainable parameters in ResNet50 were initialized with the uniform random distribution. For the ECG matrix and Mel spectrogram, the input size is 307 × 230. They are randomly cropped at 224 × 224 and flipped horizontally for better training. The network was trained by Adam optimization. The learning rate is 0.0001. The network is trained using a batch size of 32 for 80 epochs.

During validation, the input of the ECG matrix and Mel spectrogram are also center-cropped at the same size of training without horizontal flipping.

I3D with temporal attention module parameters settings

During training, the trainable parameters in I3D with the temporal attention module were also initialized with a uniform random distribution. For facial expressions, 64 consecutive frames of facial expressions are randomly sampled from each video. Input frames are rescaled to 224 * 270 and randomly cropped to 224 * 224. Frames are randomly flipped horizontally for data augmentation to improve the invariance properties of geometric perturbations. The network was trained by the Adam optimization. The learning rate is 0.01. The network is trained using batch size 1*3*64 for 30 epochs.

During validation, the input frames of facial expressions are sampled at the fixed central location of each video. These frames are rescaled and center-cropped at the same size of training without horizontal flipping.

Multimodality stress detection

The deep learning framework fused ECG, voice, and facial expressions for stress detection. Information about stress in multimodality can be obtained from the fully connected layer of ResNet50 and I3D with the temporal attention module in the framework. The framework integrated them into a global matrix for representation and used the fusion method based on matrix eigenvectors to detect stress.

The performance of the proposed deep learning framework and every single-modality-based method in the framework for stress detection was compared in the terms of four widely used metrics: accuracy, precision, recall, and F1-score. As expected, the multimodality method provided the best performance in stress detection, suggesting that the deep learning framework using multimodality data for stress detection can achieve better stress detection performance than the single-modality-based method.

As illustrated in Table 1, stress detection using multimodality improved performance compared to using only single-modality data. The accuracy of multimodality result is increased from the highest accuracy of the single-modality result of 83.9–85.1%. Revealing the deep learning framework can efficiently fuse the multimodality information for stress detection and is more effective than single-modality-based methods.

Moreover, the results demonstrated that using either single-modality or multimodality all can effectively detect stress. This work can use either of them to distinguish between calm and stress, which allows the method to be applied in situations where multimodality data are not available.

TABLE 1 Stress detection accuracy, precision, recall and F1-score using single- and multimodality data.

Modal	Accuracy	Precision	Recall	F1-Score
ECG	0.741	0.737	0.743	0.731
Voice	0.830	0.825	0.829	0.827
Facial expressions	0.792	0.795	0.803	0.799
Fusion	**0.851**	**0.857**	**0.866**	**0.861**

The bold values are the result of the multimodality stress detection.

TABLE 2 Stress detection confusion matrix of the single-modal and multimodality methods.

Actual labels	Predicted labels							
	ECG		Voice		Facial expressions		Fusion	
	Calm	Stress	Calm	Stress	Calm	Stress	Calm	Stress
Calm	0.751	0.266	0.821	0.164	0.868	0.262	0.913	0.193
Stress	0.249	0.734	0.179	0.836	0.132	0.738	0.087	0.807

Single-modality stress detection

Stress causes changes in people's involuntary body functions, and each modal of the changes can be used for stress detection. Compared with the handcrafted feature engineering methods, the deep learning network can automatically extract multiple features of the input. This work explored the use of Resnet50 and I3D with the temporal attention module to extract features in ECG, voice and facial expressions for stress detection.

After feature extraction, each single-modality was used for stress detection after simple preprocessing. The confusion matrices of each single-modal based method are shown in Table 2, and the matrices are also compared with the multimodality-based method.

In the single-modality-based method, the best recognition of the calm state is achieved by facial expressions of 86.8%, and the best recognition of the stress state is achieved by voice of 83.6%. Furthermore, the matrices also prove that ResNet50 and I3D with TAM can effectively extract stress-related features in each modality after simple preprocessing.

Effectiveness of the temporal attention module

To explore the effectiveness of TAM, an ablation experiment was designed that removed TAM for I3D with TAM. The confusion matrices of I3D without TAM are shown in Table 3. Compared with the confusion matrices of I3D with TAM in Table 2, the overall detection performance and the recognition of the calm state is improved with TAM, which proves that TAM can enhance the perception of temporal information between the calm state and stress state in I3D and emphasize the distinguishing temporal features in facial expressions.

TABLE 3 Stress detection confusion matrix of I3D without temporal attention module.

Actual labels	Predicted labels	
	Calm	Stress
Calm	0.824	0.261
Stress	0.176	0.739

This work also compared the performance of I3D without TAM and I3D with TAM in terms of the above four widely used metrics. As shown in Figure 10, without TAM, the accuracy, precision, recall, and F1-score of stress detection by facial expressions dropped by approximately 1.7, 2.0, 2.1 and 2.1%, respectively. The result demonstrates the feasibility of the TAM in finding the more influential association between frames in facial expressions about stress and I3D with TAM can achieve better performance in stress detection.

Time assessment for real-time applications

This work also analyzed the time duration of the real-time deep learning framework to verify that real-time performance requirements are met. The results show that the framework meets the needs of real-time stress assessment. Each process present in the framework was evaluated which mainly consists of three parts, namely preprocessing, feature extraction, and multimodality fusion. The results of the time duration are visualized in Figure 11.

Comparison with widely used convolutional neural networks

In this part, the comparison experiment is conducted with several widely used CNNs in each modality to evaluate the effectiveness of our work. The CNNs are ResNet101, GoogLeNet, EfficientNet, and C3D (Szegedy et al., 2015; Tran et al., 2015; He et al., 2016; Tan and Le, 2019), which are widely used and have been proven to have a strong performance. Table 4 presents their stress detection results.

As is indicated in Table 4, the accuracy of ResNet101 for stress detection by ECG and voice are 70.6 and 80.2%, both of which are lower than ResNet50. This proves that although Resnet has shortcut connections, blindly increasing the network depth cannot greatly improve the performance and will also increase the number of FLOPs. GoogLeNet and EfficientNet achieved 72.3 and 76.3%,

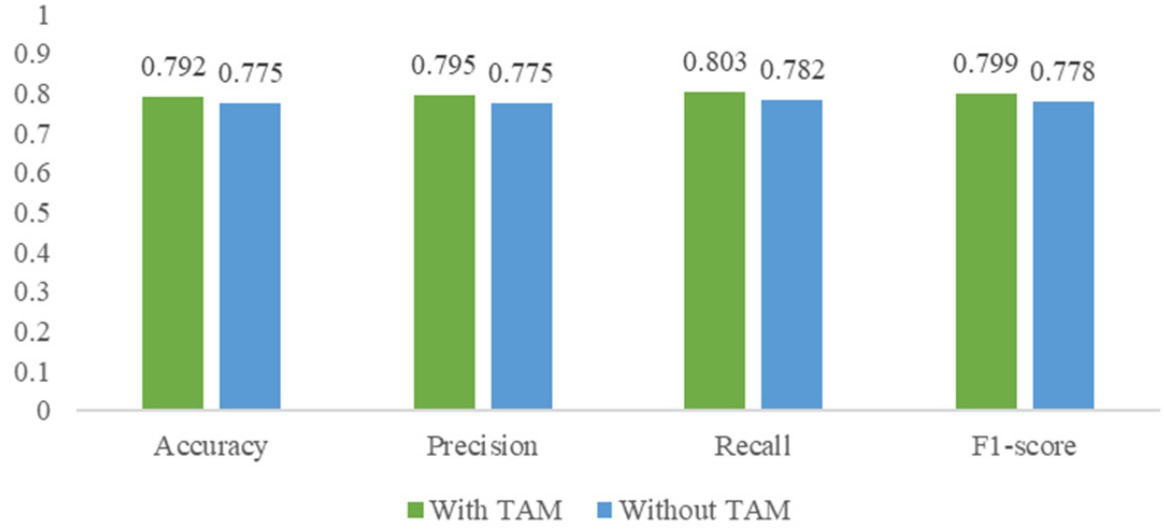

FIGURE 10
Performance comparison of I3D without temporal attention module (TAM) and I3D with TAM.

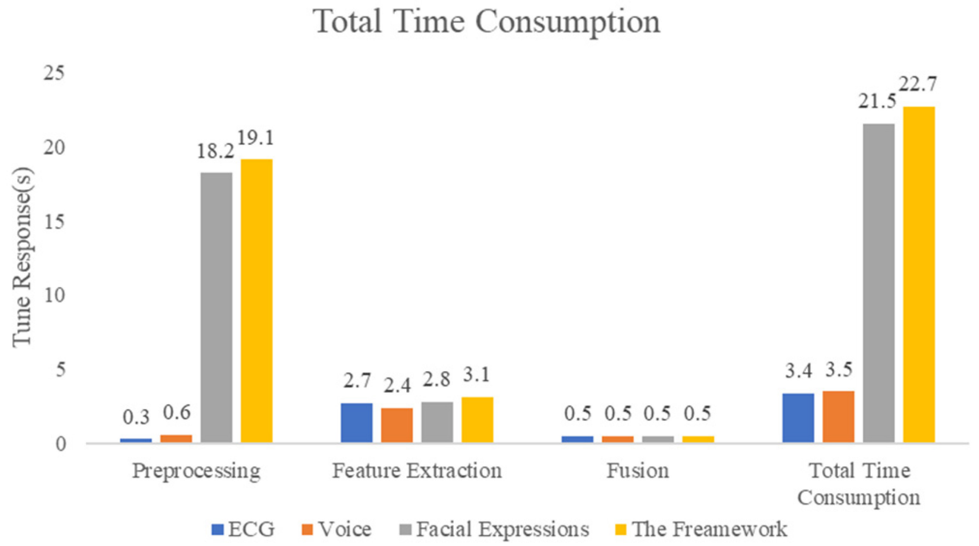

FIGURE 11
Time assessment for real-time applications.

and 76.9 and 80.9% accuracy for stress detection using ECG and voice, respectively, owing to different structures being used to solve the problems of gradient disappearance or explosion.

Table 4 also shows that I3D with TAM outperforms I3D and C3D, achieving the highest accuracy for stress detection using facial expressions. Compared with C3D, I3D can better extract the overall features of facial expressions through the inflated inception module, while TAM can highlight the distinguishing information in the overall features and find the keyframes in facial expressions to achieve optimal stress detection performance. The I3D with TAM proposed in this work can simultaneously extract the distinguishing and overall features of facial expressions through the inflated inception module and TAM. The feature maps extracted by I3D with TAM have more information, which can better use facial expressions for stress detection.

Since ResNet50 achieve 83.0% accuracy for stress detection using voice and EfficientNet achieve 76.9% accuracy for stress detection using voice. We explore EfficienNet for stress detection using ECG and voice, and I3D with TAM for stress detection using facial expressions. The same fusion method is used to fuse multimodality information for stress detection. The results are shown in Table 5.

Although EfficientNet produces good performance in stress detection using ECG and voice, the comparison of the fusion results shown in Table 5 reveals that ResNet50-based fusion method can achieve better fusion accuracy. We prove that our proposed real-time deep learning framework based on ResNet50 achieves better performance in stress detection.

Discussion

The automatic stress detection in people with objective indicators demonstrated that it can be reliable and does not require many human resources. It avoids the inference of stress detection caused by the instantaneous and subjective psychological evaluation. However, many stress detection research uses a single-modality-based for stress detection. For the multimodality data of people under acute stress, single-modality-based stress detection does not fully use all the collected data. Therefore, this work explored a way to fuse multimodality for acute stress detection.

In this work, a real-time deep learning framework was proposed to fuse multimodality for acute stress detection. Our work is trying to detect acute stress. We do not detect different levels of stress. This work proposed a fusion method based on matrix eigenvectors to fuse multimodality information. Furthermore, we designed the temporal attention module (TAM) to find the keyframes related to acute stress in facial expressions.

TABLE 4 Stress detection accuracy, precision, recall and F1-score of several widely used convolutional neural networks.

Modal	Model	Accuracy	Precision	Recall	F1-score
ECG	ResNet101[98]	0.706	0.712	0.717	0.714
	GoogLeNet[97]	0.723	0.735	0.739	0.737
	EfficientNet[99]	0.769	0.773	0.781	0.777
	ResNet50[98]	0.741	0.737	0.743	0.740
Voice	ResNet101[98]	0.802	0.797	0.801	0.799
	GoogLeNet[97]	0.763	0.756	0.752	0.754
	EfficientNet[99]	0.809	0.804	0.812	0.808
	ResNet50[98]	0.830	0.825	0.829	0.827
Facial expressions	C3D[100]	0.582	0.582	0.500	0.538
	I3D[101]	0.775	0.775	0.782	0.778
	I3D with TAM	0.792	0.795	0.803	0.799

TABLE 5 Stress detection accuracy, precision, recall and F1-score by multimodality using different convolutional neural networks.

Fusion	Accuracy	Precision	Recall	F1-score
EfficientNetI3D with TAM	0.839	0.850	0.858	0.854
ResNet50 I3D with TAM	0.851	0.857	0.866	0.861

MIST can stimulate acute stress in both the control and experimental stage by measuring changes in participants' cortisol. To evaluate the performance of the proposed framework, the multimodality dataset was collected from 20 participants during the MIST experiment. Compared to the number of participants in other stress detection research, many research collects stress data from 10-30 participants. Xia analyzed variations in both electroencephalogram (EEG) and ECG signals from 22 male subjects (Xia et al., 2018). Minguillon collected multiple biosignals from 10 subjects (Minguillon et al., 2018). Perez-Valero conducted a group of 23 participants over the MIST experiment (Perez-Valero et al., 2021). To confirm whether the participants developed acute stress during the MIST experiment, we asked each participant to fill out a questionnaire after the MIST experiment. In the questionnaires, all participants reported experiencing acute stress in both the control and experimental stages of MIST. Given the inter-individual differences in the reaction to MIST, we considered the acute stress generated in both phases as the same category. The data were labeled as "calm" and "stress" instead of different stress levels.

This work provides a method for stress detection using multimodality. The method achieves 74.1, 79.2, and 83.0% detection accuracy using ECG, facial expressions, and voice, respectively. In the single-modality-based method using facial expressions, using I3D with TAM achieves a 1.7% higher detection accuracy than using I3D. After the probability information in every single modality is fused by the proposed multimodality fusion method, the detection accuracy of 85.1% can be achieved. The results show that the overall stress detection performance is improved by using multimodality. In our results, the recognition of the stress state for the fusion method is lower than the single-modality-based method using voice. This is caused by the accuracy of the single-modality method using ECG and the accuracy of the single-modality method using voice. In our multimodality fusion method, the stress information matrix is constructed for each multimodality sample. For the recognition of the stress state, the single-modality method using ECG has an accuracy of 73.4%, and the single modality method using facial expressions has an accuracy of 73.8%. This leads to an increased probability that two or three modalities' stress information in a sample is simultaneously opposed to the real state. When the main probability information of the two modalities is opposite to the real state, it will dominate the main probability direction of the stress information matrix. This caused the stress information matrix to be more likely at the opposite of the real state, leading to incorrect recognition results. The results in this work show that the overall stress detection performance is improved by using multimodality.

Compared with other methods using other signals, a variety of objective indicators are used for stress detection. The other signals such as EEG, electromyogram (EMG), and functional near-infrared spectroscopy (fNIRS) require special equipment and a lot of pre-collection preparations and post-collection work, it limits the practical application of these signals and is unpleasant for the experimental participants. Our work demonstrates the reliability of detecting acute stress using ECG, voice, and facial expressions. The results show that using those feasible multimodality can achieve 85.1% stress detection accuracy. In addition, the modalities used in this work are easy to be acquired in daily life.

Conclusion

In this work, a real-time deep learning framework was proposed to fuse ECG, voice, and facial expressions for stress detection. The result shows that the fusion of multimodality information about stress can achieve 85.1% detection accuracy, which provides a reference for the research of multimodality stress detection based on deep learning technology in the future. The framework extracted the stress-related features of each modal through ResNet50 and I3D with TAM and gave different weights for each type of stress state according to the global stress information matrix. At the same time, this work designed the temporal attention module to find the more influential association between frames in facial expressions for stress detection. Compared with the optimal single-modality-based method, the accuracy of the multimodality result is improved by 2.1%. This work provides an objective reference for fusing multimodality to detect stress based on deep learning technology, and preventing stress from harming people's physical and mental health.

Data availability statement

The datasets presented in this article are not readily available because they contain identifiable information such as recognizable faces and must be approved by the Ethics Committee on Biomedical Research, West China Hospital of Sichuan University. Requests to access the datasets should be directed to LH, ling.he@scu.edu.cn.

Ethics statement

The studies involving human participants were reviewed and approved by the Ethics Committee on Biomedical Research, West China Hospital of Sichuan University. The patients/participants provided their written informed consent to participate in this study. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

Funding

This work was supported by the Science and Technology Department of Sichuan Province (Nos. 2020YFG0081 and 2020YFG0463-LH), Sichuan University-Yibin School-City Strategic Cooperation Special Fund Project (No. 2020CDYB-27), and State Grid Sichuan Electric Power Company (Nos. 52191918003L and 131933135027).

Acknowledgments

We would like to thank all the participants that participated in the study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aitken, R. C. (1969). A growing edge of measurement of feelings [abridged]. *Proc. R. Soc. Med.* 62, 989–993. doi: 10.1177/003591576906201005

Alberdi, A., Aztiria, A., and Basarab, A. (2016). Towards an automatic early stress recognition system for office environments based on multimodal measurements: a review. *J. Biomed. Inf.* 59, 49–75. doi: 10.1016/j.jbi.2015.11.007

Allen, A., Kennedy, P., Dockray, S., Cryan, J., Dinan, T., and Clarke, G. (2017). The trier social stress test: principles and practice. *Neurobiol. Stress* 6, 113–126. doi: 10.1016/j.ynstr.2016.11.001

Beanland, V., Fitzharris, M., Young, K., and Lenné, M. (2013). Driver inattention and driver distraction in serious casualty crashes: data from the australian national crash in-depth study. *Accid. Anal. Prev.* 54, 99–107. doi: 10.1016/j.aap.2012.12.043

Boehringer, A., Tost, H., Haddad, L., Lederbogen, F., Wüst, S., Schwarz, E., et al. (2015). Neural correlates of the cortisol awakening response in humans. *Neuropsychopharmacology* 40, 2278–2285. doi: 10.1038/npp.2015.77

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, (Santiago), 6299–6308.

Chung, K. C., Springer, I., Kogler, L., Turetsky, B., Freiherr, J., and Derntl, B. (2016). The influence of androstadienone during psychosocial stress is modulated by gender, trait anxiety and subjective stress: an fMRI study. *Psychoneuroendocrinology* 68, 126–139. doi: 10.1016/j.psyneuen.2016.02.026

Cinaz, B., Arnrich, B., La Marca, R., and Tröster, G. (2011). Monitoring of mental workload levels during an everyday life office-work scenario. *Pers. Ubiquit. Comput.* 17, 229–239. doi: 10.1007/s00779-011-0466-1

De Rosa, G. (2004). Moderate physical exercise increases cardiac autonomic nervous system activity in children with low heart rate variability. *Childs Nerv. Syst.* 20, 215–215. doi: 10.1007/s00381-004-0916-4

de Santos Sierra, A., Ávila, C. S., Del Pozo, G. B., and Casanova, J. G. (2011). "Stress detection by means of stress physiological template," in *Proceedings of the 2011 Third World Congress on Nature and Biologically Inspired Computing*, (Salamanca: IEEE), 131–136. doi: 10.1109/nabic.2011.6089448

Dedovic, K., Renwick, R., Mahani, N. K., Engert, V., Lupien, S. J., and Pruessner, J. C. (2005). The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *J. Psychiatry Neurosci.* 30, 319–325.

Freeman, H. J. (1986). Environmental stress and psychiatric disorder. *Stress Med.* 2, 291–299. doi: 10.1002/smi.2460020404

Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P., et al. (2017). Stress and anxiety detection using facial cues from videos. *Biomed. Signal Process. Control* 31, 89–101. doi: 10.1016/j.bspc.2016.06.020

Gossett, E., Wheelock, M., Goodman, A., Orem, T., Harnett, N., Wood, K., et al. (2018). Anticipatory stress associated with functional magnetic resonance imaging: implications for psychosocial stress research. *Int. J. Psychophysiol.* 125, 35–41. doi: 10.1016/j.ijpsycho.2018.02.005

Hakimi, N., and Setarehdan, S. K. (2018). Stress assessment by means of heart rate derived from functional near-infrared spectroscopy. *J. Biomed. Opt.* 23:115001. doi: 10.1117/1.jbo.23.11.115001

Hatcher, W., and Yu, W. (2018). A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access* 6, 24411–24432. doi: 10.1109/access.2018.2830661

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, (Las Vegas, NV), 770–778. doi: 10.1109/cvpr.2016.90

Heraclides, A., Chandola, T., Witte, D., and Brunner, E. (2012). Work stress, obesity and the risk of type 2 diabetes: gender-specific bidirectional effect in the whitehall II study. *Obesity* 20, 428–433. doi: 10.1038/oby.2011.95

Hwang, B., You, J., Vaessen, T., Myin-Germeys, I., Park, C., and Zhang, B. (2018). Deep ECGNet: an optimal deep learning framework for monitoring mental stress using ultra short-term ECG signals. *Telemed. J. E Health* 24, 753–772. doi: 10.1089/tmj.2017.0250

Jin, L., Xue, Y., Li, Q., and Feng, L. (2016). "Integrating human mobility and social media for adolescent psychological stress detection," in *Proceedings of the International Conference on Database Systems for Advanced Applications*, (Cham: Springer), 367–382. doi: 10.1007/978-3-319-32049-6_23

Kiem, S., Andrade, K., Spoormaker, V., Holsboer, F., Czisch, M., and Sämann, P. (2013). Resting state functional MRI connectivity predicts hypothalamus-pituitary-axis status in healthy males. *Psychoneuroendocrinology* 38, 1338–1348. doi: 10.1016/j.psyneuen.2012.11.021

Kirschbaum, C., Pirke, K., and Hellhammer, D. (1993). The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 76–81. doi: 10.1159/000119004

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lederbogen, F., Kirsch, P., Haddad, L., Streit, F., Tost, H., Schuch, P., et al. (2011). City living and urban upbringing affect neural social stress processing in humans. *Nature* 474, 498–501. doi: 10.3410/f.13813956.15250056

Li, F., Xu, P., Zheng, S., Chen, W., Yan, Y., Lu, S., et al. (2018). Photoplethysmography based psychological stress detection with pulse rate variability feature differences and elastic net. *Int. J. Distributed Sens. Netw.* 14:1550147718803298. doi: 10.1177/1550147718803298

Liao, W., Zhang, W., Zhu, Z., and Ji, Q. (2005). "A real-time human stress monitoring system using dynamic bayesian network," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, (San Diego, CA: *IEEE*), 70. doi: 10.1109/cvpr.2005.394

Lovallo, W. (1975). The cold pressor test and autonomic function: a review and integration. *Psychophysiology* 12, 268–282. doi: 10.1111/j.1469-8986.1975.tb01289.x

Lu, H., Frauendorfer, D., Rabbi, M., Mast, M. S., Chittaranjan, G. T., Campbell, A. T., et al. (2012). "Stresssense: detecting stress in unconstrained acoustic environments using smartphones," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, (Pittsburgh, PA), 351–360. doi: 10.1145/2370216.2370270

McDuff, D., Karlson, A., Kapoor, A., Roseway, A., and Czerwinski, M. (2012). "AffectAura: emotional wellbeing reflection system," in *Proceedings of the 2012 6th international conference on pervasive computing technologies for healthcare (PervasiveHealth) and workshops* (Piscataway, NY: IEEE), 199–200. doi: 10.4108/icst.pervasivehealth.2012.248727

Minguillon, J., Perez, E., Lopez-Gordo, M., Pelayo, F., and Sanchez-Carrion, M. (2018). Portable system for real-time detection of stress level. *Sensors* 18:2504. doi: 10.3390/s18082504

Mitra, A. (2008). Diabetes and stress: a review. *Stud. Ethnomed.* 02, 131–135. doi: 10.31901/24566772.2008/02.02.07

Nair, V., and Hinton G. E. (2010). "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th international conference on international conference on machine learning (ICML'10)*, Haifa, 807–814. doi: 10.5555/3104322.3104425

Nigam, K., Godani, K., Sharma, D., and Jain, S. (2021). An improved approach for stress detection using physiological signals. *ICST Trasnsac. Scalable Inf. Syst.* 8:169919. doi: 10.4108/eai.14-5-2021.169919

Niu, L., Chen, C., Liu, H., Zhou, S., and Shu, M. (2020). A deep-learning approach to ECG classification based on adversarial domain adaptation. *Healthcare* 8:437. doi: 10.3390/healthcare8040437

Noack, H., Nolte, L., Nieratschker, V., Habel, U., and Derntl, B. (2019). Imaging stress: an overview of stress induction methods in the MR scanner. *J. Neural Transm.* 126, 1187–1202. doi: 10.1007/s00702-018-01965-y

Pampouchidou, A., Pediaditis, M., Chiarugi, F., Marias, K., Simos, P., Yang, F., et al. (2016). "Automated characterization of mouth activity for stress and anxiety assessment," in *Proceedings of the 2016 IEEE International Conference on Imaging Systems and Techniques (IST)* (Chania), 356–361. doi: 10.1109/ist.2016.7738251

Perez-Valero, E., Vaquero-Blasco, M. A., Lopez-Gordo, M. A., and Morillas, C. (2021). Quantitative assessment of stress through EEG during a virtual reality stress-relax session. *Front. Comput. Neurosci.* 15:684423. doi: 10.3389/fncom.2021.684423

Piotrowski, Z., and Różanowski, K. (2010). Robust algorithm for Heart Rate (HR) detection and Heart Rate Variability (HRV) estimation. *Acta Phys. Polonica A* 118, 131–135. doi: 10.12693/aphyspola.118.131

Reinhardt, T., Schmahl, C., Wüst, S., and Bohus, M. (2012). Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). *Psychiatry Res.* 198, 106–111. doi: 10.1016/j.psychres.2011.12.009

Ren, P., Ma, X., Lai, W., Zhang, M., Liu, S., Wang, Y., et al. (2019). Comparison of the use of blink rate and blink rate variability for mental state recognition. *IEEE Transac. Neural Syst. Rehabil. Eng.* 27, 867–875. doi: 10.1109/tnsre.2019.2906371

Renaud, P., and Blondin, J. (1997). The stress of Stroop performance: physiological and emotional responses to color–word interference, task pacing, and pacing speed. *Int. J. Psychophysiol.* 27, 87–97. doi: 10.1016/s0167-8760(97)00049-4

Sauter, S., Murphy, L., and Hurrell, J. (1990). Prevention of work-related psychological disorders: a national strategy proposed by the National Institute for Occupational Safety and Health (NIOSH). *Am. Psychol.* 45, 1146–1158. doi: 10.1037/10108-002

Segerstrom, S., and Miller, G. (2004). Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry. *Psychol. Bull.* 130, 601–630. doi: 10.1037/0033-2909.130.4.601

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Troster, G., and Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transac. Inf. Technol. Biomed.* 14, 410–417. doi: 10.1109/titb.2009.2036164

Sharma, N., and Gedeon, T. J. (2012). Objective measures, sensors and computational techniques for stress recognition and classification: a survey. *Comput. Methods Programs Biomed.* 108, 1287–1301. doi: 10.1016/j.cmpb.2012.07.003

Sharma, S., Singh, G., and Sharma, M. (2021). A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans. *Comput. Biol. Med.* 134:104450. doi: 10.1016/j.compbiomed.2021.104450

Sioni, R., and Chittaro, L. (2015). Stress detection using physiological sensors. *Computer* 48, 26–33. doi: 10.1109/mc.2015.316

Smeets, T., Cornelisse, S., Quaedflieg, C., Meyer, T., Jelicic, M., and Merckelbach, H. (2012). Introducing the Maastricht Acute Stress Test (MAST): a quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. *Psychoneuroendocrinology* 37, 1998–2008. doi: 10.1016/j.psyneuen.2012.04.012

Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., et al. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ Digit. Med.* 1:67. doi: 10.1038/s41746-018-0074-9

Stepanovic, S., Mozgovoy, V., and Mettler, T. (2019). "Designing visualizations for workplace stress management: results of a pilot study at a swiss municipality," in *Proceedings of the International Conference on Electronic Government*, (Cham: Springer), 94–104. doi: 10.1007/978-3-030-27325-5_8

Steptoe, A., and Kivimäki, M. (2013). Stress and cardiovascular disease: an update on current knowledge. *Annu. Rev. Public Health* 34, 337–354. doi: 10.1146/annurev-publhealth-031912-114452

Sundelin, T., Lekander, M., Kecklund, G., Van Someren, E., Olsson, A., and Axelsson, J. (2013). Cues of fatigue: effects of sleep deprivation on facial appearance. *Sleep* 36, 1355–1360. doi: 10.5665/sleep.2964

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, MA), 1–9. doi: 10.1109/cvpr.2015.7298594

Tan, M., and Le, Q. (2019). "Efficientnet: rethinking model scaling for convolutional neural networks," in *Proceedings of the International Conference on Machine Learning. PMLR*, (Long Beach), 6105–6114. doi: 10.1007/978-1-4842-6168-2_10

Tanosoto, T., Arima, T., Tomonaga, A., Ohata, N., and Svensson, P. (2012). A paced auditory serial addition task evokes stress and differential effects on masseter-muscle activity and haemodynamics. *Eur. J. Oral Sci.* 120, 363–367. doi: 10.1111/j.1600-0722.2012.00973.x

Tomova, L., Majdandžić, J., Hummer, A., Windischberger, C., Heinrichs, M., and Lamm, C. (2016). Increased neural responses to empathy for pain might explain how acute stress increases prosociality. *Soc. Cogn. Affect. Neurosci.* 12, 401–408. doi: 10.1093/scan/nsw146

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, (Santiago), 4489–4497.

Tsigos, C., and Chrousos, G. (2002). Hypothalamic–pituitary–adrenal axis, neuroendocrine factors and stress. *J. Psychosom. Res.* 53, 865–871. doi: 10.1016/s0022-3999(02)00429-4

van Praag, H. (2004). Can stress cause depression? *Progr. Neuropsychopharmacol. Biol. Psychiatry* 28, 891–907. doi: 10.1016/j.pnpbp.2004.05.031

Vidya, K., Ng, E., Acharya, U., Chou, S., Tan, R., and Ghista, D. (2015). Computer-aided diagnosis of myocardial infarction using ultrasound images with DWT, GLCM and HOS methods: a comparative study. *Comput. Biol. Med.* 62, 86–93. doi: 10.1016/j.compbiomed.2015.03.033

Wei, C. (2013). Stress emotion recognition based on RSP and EMG signals. *Adv. Mater. Res.* 709, 827–831.

Wheelock, M. D., Harnett, N. G., Wood, K. H., Orem, T. R., Granger, D. A., Mrug, S., et al. (2016). Prefrontal cortex activity is associated with biobehavioral components of the stress response. *Front. Hum. Neurosci.* 10:583. doi: 10.3389/fnhum.2016.00583

Wiegel, C., Sattler, S., Göritz, A., and Diewald, M. (2015). Work-related stress and cognitive enhancement among university teachers. *Anxiety Stress Coping* 29, 100–117. doi: 10.1080/10615806.2015.1025764

Winata, G. I., Kampman, O. P., and Fung, P. (2018). "Attention-based lstm for psychological stress detection from spoken language using distant supervision," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Calgary, AB: IEEE), 6204–6208. doi: 10.1109/icassp.2018.8461990

Xia, L., Malik, A. S., and Subhani, A. R. (2018). A physiological signal-based method for early mental-stress detection. *Biomed. Signal Process. Control* 46, 18–32. doi: 10.1201/9780429196621-13

Yang, Q., Li, L., Zhang, J., Shao, G., Zhang, C., and Zheng, B. (2013). Computer-aided diagnosis of breast DCE-MRI images using bilateral asymmetry of contrast enhancement between two breasts. *J. Digit. Imaging* 27, 152–160. doi: 10.1007/s10278-013-9617-4

Zhu, Q., Xu, X., Yuan, N., Zhang, Z., Guan, D., Huang, S., et al. (2020). Latent correlation embedded discriminative multi-modal data fusion. *Signal Process.* 171:107466. doi: 10.1016/j.sigpro.2020.107466