Trial Selection Tensor Canonical Correlation Analysis (TSTCCA) for Depression Recognition with Facial Expression and Pupil Diameter

Minqiang Yang, Yushan Wu, Yongfeng Tao, Xiping Hu*, and Bin Hu*

***Abstract*— Facial expressions have been widely used for depression recognition because it is intuitive and convenient to access. Pupil diameter contains rich emotional information that is already reflected in facial video streams. However, the spatiotemporal correlation between pupillary changes and facial behavior changes induced by emotional stimuli has not been explored in existing studies. This paper presents a novel multimodal fusion algorithm - Trial Selection Tensor Canonical Correlation Analysis (TSTCCA) to optimize the feature space and build a more robust depression recognition model, which innovatively combines the spatiotemporal relevance and complementarity between facial expression and pupil diameter features. TSTCCA explores the interaction between trials and obtains an effective fusion representation of two modalities from a trial subset related to depression. The experimental results show that TSTCCA achieves the highest accuracy of 78.81% with the subset of 25 trials.**

***Index Terms*— Depression recognition, Multimodal fusion, Facial expression, Pupil diameter**

I. Introduction

Depression is one of the most common mental disorders worldwide, which imposes an enormous social and economic burden [1]. Depression may seriously interfere with the patient's daily life, even leading to suicidal tendencies [2] [3]. Currently, the diagnosis of depression is primarily based on self-rating scales and clinical interviews. However, these diagnostic methods rely heavily on the patient's subjective perception and the doctor's professional knowledge [4]. Studies on disease surveillance based on physiological and behavioral data are widely conducted [5] [6], and have provided quite a few clues for the objective auxiliary diagnosis of depression [7] [8] [9].

The ability to recognize emotions is impaired in many people with mental disorders [10]. As a behavioral signal, facial expression plays a crucial role in emotion recognition [11]. Therefore, facial expression has become a relatively effective tool for identifying emotional disorders [12]. Moreover, pupil dilation has also been shown to be associated with emotion [13]. Multimodal data can capture complementary information not visible in unimodal data [14]. Literature well established the multimodal fusion method is superior to the unimodal method [15]. Many studies have extracted depression discriminative cues from multiple modalities. However, we have not found any multimodal fusion research based on facial expression and pupil diameter. Hence, this paper attempts to fuse facial expression and pupil diameter to explore a more precise depression recognition method.

Currently, feature fusion methods are divided into deep learning and traditional methods. Some representative deep learning models have been applied to multimodal fusion analysis, such as Deep Belief Net (DBN) [16], Recurrent Neural Network (RNN) [17], Autoencoder (AE) [18] [19], Convolutional Neural Network (CNN) [20], Transformer [21], etc. Deep learning often requires a large number of samples [22], but our depression assessment research has limited data. [23] used multimodal information for data enhancement. Nevertheless, it is still difficult to avoid the problem that the model is prone to overfitting. In this study, we choose the traditional fusion methods based on Canonical Correlation Analysis (CCA) [24] to accommodate the limited data set [25] [26]. CCA is widely applied in affective computing, and it can detect patterns of common variation in multimodal, which allows it not only to fuse information but also to extract the relevant features. CCA takes the effective discriminant features, which can reveal potentially significant variations in facial expression and pupil diameter [27]. Despite the profound theoretical foundation and practical success of CCA in multimodal fusion, it can only handle vector features. However, the features utilized in many real-world applications are usually multi-dimensional arrays. High-dimensional features and small-scale data sets may lead to the singularity problem of the covariance matrix of CCA [28]. To solve this problem [29], [30] proposed Two-Dimensional CCA (2DCCA), which extended CCA to matrix-valued data. 2DCCA retains a matrix representation of data but can't be applied to tensor-valued data. A typical approach for generalizing 2DCCA to tensor is to reshape each data tensor into a matrix. This strategy breaks the structure of the tensor data and leads to a curse of dimensionality [31]. [32] proposed Tensor Canonical Correlation Analysis (TCCA) to overcome this difficulty. [32] interpreted 2DCCA as a non-convex method for the low-rank tensor factorization problem, allowing them to receive a tensor extension of 2DCCA.

Although the above methods can carry out effective multimodal fusion, they can not analyze the trials in the paradigm most associated with depression. In the specific application of depression recognition, previous studies rely on the arousal and value index to select the trials and ignore the interaction between trials. Redundant features obtained using low-correlation trials can adversely affect true features and produce heterogeneity in the data [33]. Selecting a trial suitable for depression recognition can also reduce feature dimensionality to eliminate the effects of noise, which helps classifiers focus on essential features and ignore misleading features [34]. Moreover, the optimal subset of trials can decrease the computational complexity of experimental data and help avoid over-fitting, leading to a better performance of depression recognition [8].

For integrating the complementary information from two feature tensors, we propose Trial Selection Tensor Canonical Correlation Analysis (TSTCCA) to combine facial expression and pupil diameter for depression recognition. Mutual information is used to measure the spatiotemporal correlation between two modalities trials. Intending to increase the reliability of estimates, TSTCCA selects the trial by the mutual information between feature tensors of two modalities and explores optimal representations from multimodal tensor data sets, which is also the origin of the name TSTCCA. The Support Vector Machines (SVM), K-Nearest Neighbor (kNN), and Random Forest (RF) are utilized to detect depression. The main contributions of this paper include the following three points.

Minqiang Yang, Yushan Wu, and Yongfeng Tao are with School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China(e-mail: yangmq@lzu.edu.cn, wuysh2021@lzu.edu.cn, taoyf21@lzu.edu.cn).

Xiping Hu is with Beijing Institute of Technology, Beijing 100081, China(e-mail: huxp@bit.edu.cn).

Bin Hu is with Gansu Provincial Key Laboratory of Wearable Computing, School of Information Science and Engineering, Lanzhou University, Gansu, China. He is also with School of Medical Technology,Beijing Institute of Technology, Beijing, China(e-mail: bh@lzu.edu.cn).

*Corresponding author

- It proposes a new multimodal feature fusion method TSTCCA for depression recognition. TSTCCA measures the spatiotemporal correlation and complementarity between facial expression and pupil diameter in tensor feature space and explores optimal representations from multimodal tensor data sets.
- It conducts a pioneering study of multimodal fusion research based on facial expression and pupil diameter. The experimental results verify the validity of the complementarity of facial behavioral phenotypes and pupil physiological characteristics in depression assessment.
- It discusses preferred trial subsets related to depression based on the retained trials, which reveals that negative and positive trials are always more likely to be selected. It also provides an opportunity for paradigm optimization that may help support the diagnosis of depression in the future.

The remainder of this paper is organized as follows. Section II discusses the related work of the current study. In Section III, we introduce the design of our experiments. In Section IV, the TSTCCA method is presented in detail. Section V describes the depression recognition experiments and classification performance. Then, the experimental results are discussed in Section VI. Finally, Section VII concludes the paper.

II. Related work

Facial expressions provide discriminative cues for depression analysis. Depressed individuals present low expressibility of facial expressions [35] [36]. The "Audio/Visual Emotion Challenge" (AVEC) series inspired extensive video and audio depression recognition research [37]. In image modality, [38] presented a deep architecture termed DepressNet to recognize depression through visual representations. [39] combined 2D-CNN networks and distributed learning to estimate depression levels. The above studies used deep architectures to pre-train their models and fine-tune their models through AVEC database [40]. To improve the accuracy of depression recognition, some scholars trained deep models from scratch rather than pre-training. [41] extracted human behavior primitives features (action units, gaze direction, and head pose) from the video and fed them to a multi-scale network for depression analysis. [42] integrated the attention mechanism into the 2D-CNN network and proposed a novel architecture named Deep LocalGlobal Attention Convolutional Neural Network (DLGA-CNN). Although the single image contained a lot of discriminant information related to depression recognition, the temporal information in the video was neglected. [43] presented an end-to-end intelligent system to extract the representation of the entire video clip. [44] proposed the Maximization and Differentiation Network (MDN) to address the overfitting problem of 3D-CNN. [45] presented a depression recognition framework in the 5G mobile network scenario and conducted the validation experiment.

Eye movement is an essential behavioral signal, which has been widely used in depression recognition [46] [47]. [48] extracted eleven eye movement features from three tests (fixation stability, free-viewing, and anti-saccade tests) to classify 65 volunteers, and the model achieved an accuracy of 86.0%. [49] combined eye movements and the attentional bias theory and obtained psychological features from eye movement data. The accuracy reached 77.0% when using the SVM classifier. [9] assessed the application of face and eye movement tracking during cognitive task performance for depression recognition.

As one of the many eye movement indicators, pupil diameter can be well measured automatically [50]. [51] investigated the classification abilities of several eye movement features for five human emotions. The conclusion was that pupil diameter had a higher discrimination ability for affective classification than the other eye movement features. [52] showed that the pupillary motions of depressed patients were different from normal controls. In [53], it is considered that depressed individuals displayed more intense sustained pupil dilation than never-depressed ones after emotional stimuli. [54] analyzed the relationship between motivational states and affective processes and found depressed participants with more highly motivated had more significant pupillary responses. [55] proved that the pupillary response to light in depressed patients significantly differed from that in normal controls. [56] also revealed that pupil diameter was a significant indicator of depression assessment.

Multimodal fusion method can represent different modalities uniformly, whose ability outperforms the unimodal method [15]. More and more scholars fused data from different distributions, sources, and types to enhance the performance of depression recognition. Based on AVEC database, [57] presented a multimodal spatiotemporal representation architecture to predict the severity of depression. The multimodal attention feature fusion (MAFF) method and the spatiotemporal attention (STA) network make outstanding contributions to the proposed architecture. [58] adopted facial expression, movement, Self-Reported Depression Scale (SDS) information, and Self-Reported Anxiety Scale (SAS) information to obtain a diagnostic framework for depression, which consisted of CNN and long short-term memory (LSTM) network. [59] fused electroencephalography (EEG), pupil diameter, and other eye movement features (blink times, saccade counts, etc.) at the decision level, and proposed a content-based multiple evidence fusion (CBMEF) method. In [19], pupil area signals and EEG are fused through the denoising autoencoder and built Mutual Information Based Fusion Model (MIBFM) for mild depression detection.

In recent years, though most research on multimodal fusion has been based on deep learning, some scholars attempted to apply traditional methods. CCA is a powerful multimodal fusion method that embeds two sets of variables with different dimensions into complex high-dimensional spaces [60]. [61] utilized CCA to fuse structural magnetic resonance imaging (sMRI), functional magnetic resonance imaging (fMRI), and electroencephalogram (EEG) data of people with schizophrenia. [62] used CCA to uncover patterns related to mild cognitive impairment by applying it to sMRI and fMRI data. [63] achieved diagnosing Alzheimer's disease using the fusion method based on CCA. After decades of development, scholars have proposed various improved algorithms based on CCA, and these methods also have excellent performance in the field of multimodal fusion. 2DCCA implemented the analysis of data without damaging the 2D feature structure. To operate 2D fMRI images directly, [64] introduced 2DCCA into multiset canonical correlation analysis (MCCA) structure for multi-subject medical images analysis. [65] applied 2DCCA to fuse high-resolution multispectral (HRM) images through the low-resolution multispectral (LRM) and high-resolution panchromatic (HRP) images. [66] employed deep canonical correlation analysis (DCCA) to integrate multimodal emotion data into a hyperspace.

Although many studies have focused on recognizing depression through multimodal data, few have fused facial expressions and pupil diameters, and previous studies ignored the interaction between trials. The proposed TSTCCA selects the trial by the dependence between facial expression and pupil diameter feature tensors and explores optimal representations from multimodal tensor data sets, which provides a novel method for multimodal depression recognition.

III. Paradigm and Data

TABLE I: Stimulus paradigm.

Continuous Stimulus																										
			Block1																							
	Segment	Focus	Neutral					Focus	Positive					Focus	Neutral					Focus	Negative					Focus
	Identifier	' + '	1	2	3	4	5	' + '	6	7	8	9	10	' + '	11	12	13	14	15	' + '	16	17	18	19	20	' + '
	Trial		room	basket	outlet	clock	clothespins		puppies	family	seal	butterfly	athletes		chair	coffee cup	key ring	book	abstract art		mutilation	toilet	burn	shadow	hand	
	Duration Time(s)	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
			Block2																							
	Segment		Neutral					Focus	Positive					Focus	Neutral					Focus	Negative					Focus
	Identifier		21	22	23	24	25	' + '	26	27	28	29	30	' + '	31	32	33	34	35	' + '	36	37	38	39	40	' + '
	Trial		plate	abstract art	bowl	clock	tool		giraffes	bunnies	women	mother	adult		cabinet	spoon	light bulb	mug	mug		snake	victim	snakes	baby	sick kitty	
	Duration Time(s)		5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

A. Paradigm

The International Affective Picture System (IAPS) is the most used standardized emotion elicitation database [67], which has been widely used in affective computing research [68] [69]. IAPS can provide robust emotional induction in a controlled environment, even for different subjects. Therefore, we select 10 positive, 20 neutral, and 10 negative emotional pictures from IAPS to design a stimulus paradigm with emotional effects for depression recognition. The experiment contains two blocks, and each block consists of four segments. Each segment includes the annotation focus and five stimulation pictures. Each stimuli picture is considered a trial, and more details are shown in Table 1.

During the experiment, subjects sit in front of a computer screen at about 50-70 cm distance. Under the guidance of the experimenter, the subjects complete the calibration and watch the displayed trials sequence freely. Each trial plays for 5 seconds, and each rest period between two sets of trials plays for 5 seconds. We also asked subjects to reduce body and head movements during the experiment.

Facial expressions are collected by Logitech C1000, the device has 30 fps frame rate and 1920 $\times$ 1080 P resolution. Since the desktop eye tracker collects the whole body image, the algorithm needs to crop the eye image, resulting in a low resolution of the eye image, and the existing wearable will occlude some crucial facial landmarks. Hence, an eye tracker is redesigned to record pupil changes. The frame rate of eye tracking is 200 fps, and the resolution is 320 $\times$ 200 P.

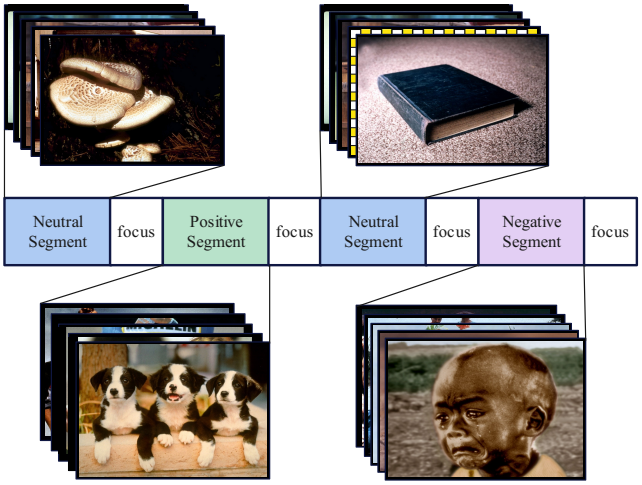


Fig. 1: The paradigm process.

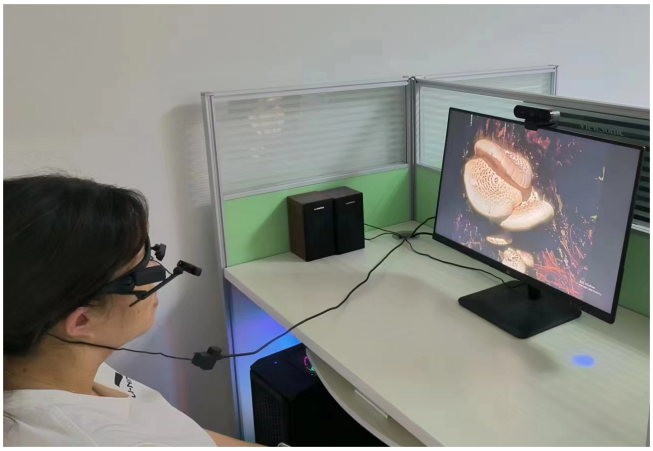


Fig. 2: The experimental environment.

B. Subjects

This study is approved by the ethics committee of the Third People's Hospital of Tianshui City. Before the experiment, all subjects signed the written informed consent. The ages of all subjects are between 18 and 55 years old [70], and all subjects at least have a primary school education level. All subjects haven't got psychotropic drug treatment in the last two weeks, and have no severe suicidal tendencies or serious physical illnesses. The normal controls have no personal or family history of mental illness. After receiving the structured Mini International Neuropsychiatric Interview (M.I.N.I.) and the Patient Health Questionnaire (PHQ-9), all the depressed patients meet the DSM-IV major depression diagnostic criteria and the PHQ-9 score $\geq$5. We use 106 valid data, including 53 depressed subjects (15 males and 38 females) and 53 normal control subjects (14 males and 39 females). Since the gender ratio between the depression group and the comparison group is basically balanced, the effect of gender is ignored in the data analysis process.

IV. Method

In order to explore the effects of different emotional stimuli on patients with depression and realize effective recognition of depression, we propose TSTCCA to fuse facial expression and pupil diameter features. Compared with CCA and TCCA, TSTCCA explores the spatiotemporal correlation between two modalities, which not only eliminates the low-correlation trials to construct a suitable paradigm that better reflects the differences between the depressed patient and normal control, but also optimizes the feature representation to reduce the computational complexity.

The TSTCCA flow chart is shown in Fig. 3. The framework in the figure mainly comprises preprocessing, feature extraction, mutual information calculation, and fusion classification modeling. First, the raw facial video is preprocessed to identify facial regions by face alignment, and the time series of pupil diameters are generated by the software shipped with the eye tracker and denoised. Histogram of Oriented Gradient (HOG) features are extracted for facial video, while multidimensional time domain and frequency domain features are extracted for pupil time series. Mutual information is used to analyze and measure the dependency between the two modalities'

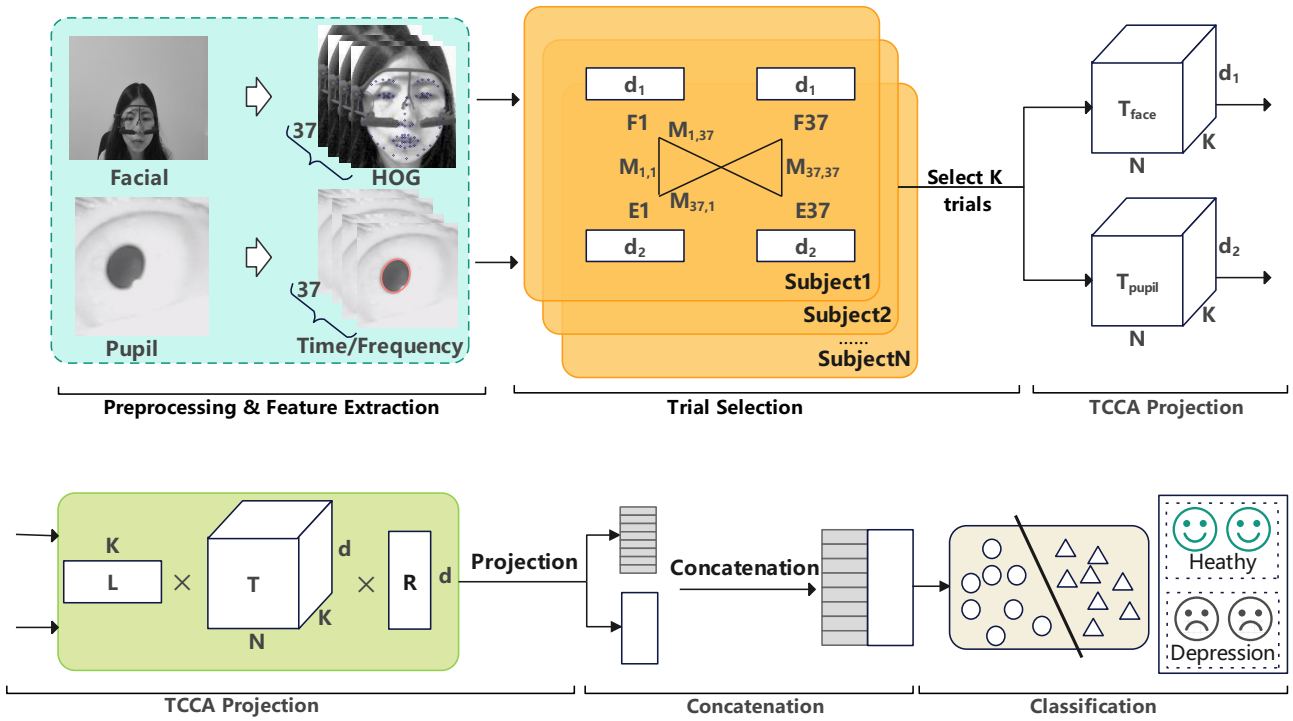


Fig. 3: The flow chart of TSTCCA.

features, based on which trial selection is conducted. We analyze the correlation between two modalities' feature tensors and project the tensors as vectors. Then, two projected feature vectors are horizontally concatenated for feature layer fusion of TCCA. Finally, we use SVM, kNN, and RF algorithms to classify the bimodal fused representations.

In section A, we first provide the necessary notation and concepts for algebraic operation. Then we introduce CCA and TCCA methods in section B. Finally, the TSTCCA method is described in detail in section C.

A. Notation and Concepts in Algebraic Operation

Before introducing the methods, some notation and concepts in algebraic operation must be defined. Scalars, vectors, matrices, and higher-order tensors are denoted by lower-case letters $(a, b, c, \cdots)$, upper-case letters $(A, B, C, \cdots)$, bold capital letters $(\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \cdots)$, and calligraphic letters $(\mathcal{A}, \mathcal{B}, \mathcal{C} \cdots)$, respectively. Let $\mathcal{X}$ and $\mathcal{Y}$ be n-order tensors of size $d_1 \times d_2 \cdots \times d_n$. The inner product on $\mathcal{X}$ and $\mathcal{Y}$ is defined as

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i_1}^{d_1} \cdots \sum_{i_n}^{d_n} x_{i_1 i_2 \cdots i_n} y_{i_1 i_2 \cdots i_n} \tag{1}$$

The mode-k product of $\mathcal{X}$ with a matrix $\boldsymbol{G}$ of size $g \times d_k$ is a tensor of size $d_1 \times \cdots \times d_{k1} \times g \times d_{k+1} \cdots \times d_n$ given by

$$(\mathcal{X} \times_k G)_{i_1 \cdots i_{m-1} j i_{m+1} \cdots i_n} = \sum_{i_k=1}^{d_k} x_{i_1 i_2 \cdots i_n} g_{j i k} \tag{2}$$

Let A_1, $A_2 \cdots A_n$ be vectors of size d_1, $d_2 \cdots d_n$, respectively. The outer product of A_1, $A_2 \cdots A_n$ is an n-order tensor defined by

$$(A_1 \circ \cdots \circ A_n)_{i_1 i_2 \cdots i_n} = (A_1)_{i_1} \cdots (A_n)_{i_n} \tag{3}$$

The Kronecker product of two matrices $\boldsymbol{E}$ of size $m \times n$ and $\boldsymbol{F}$ of size $p \times q$ is an $mp \times nq$ matrix defined by

$$E \otimes F = (e_{ij} F)_{mp \times nq} \tag{4}$$

If there are vectors $Y_1, \cdots, Y_d$ and $\mathcal{Y} = Y_1 \circ \cdots \circ Y_d$, $\mathcal{Y}$ is a rank-one tensor.

B. Related Methods

Suppose there are two matrices $\boldsymbol{X} \in \mathbb{R}^{n \times d_1}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times d_2}$. The basic idea of CCA is to seek two sets of projection vectors $A \in \mathbb{R}^{d_1 \times 1}$ and $B \in \mathbb{R}^{d_2 \times 1}$ to maximize the correlation between the canonical variables $X' = \boldsymbol{X}A$ and $Y' = \boldsymbol{Y}B$.

$$\begin{aligned} &\arg\max_{A,B} A^T \boldsymbol{C_{XY}} B \\ &\text{s.t. } A^T \boldsymbol{C_{XX}} A = 1,\ B^T \boldsymbol{C_{YY}} B = 1 \end{aligned} \tag{5}$$

where $\boldsymbol{C_{XY}}$ corresponds to the cross-covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{C_{XX}}$ and $\boldsymbol{C_{YY}}$ correspond to the auto-covariance of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Eq. (5) can be solved by the Singular Value Decomposition (SVD) after matrix standardization. The optimization objective becomes the following equation

$$\begin{aligned} &\arg\max_{A,B} U^T \boldsymbol{C_{XX}}^{-\frac{1}{2}} \boldsymbol{C_{XY}} \boldsymbol{C_{YY}}^{-\frac{1}{2}} V \\ &\text{s.t. } U^T U = 1,\ V^T V = 1 \end{aligned} \tag{6}$$

where $U = \boldsymbol{C}_{\boldsymbol{XX}}^{\frac{1}{2}} A$ and $V = \boldsymbol{C}_{\boldsymbol{XX}}^{\frac{1}{2}} B$. Suppose U and V are a pair of left and right singular vectors of $\boldsymbol{M} = \boldsymbol{C}_{\boldsymbol{XX}}^{-\frac{1}{2}} \boldsymbol{C_{XY}} \boldsymbol{C}_{\boldsymbol{YY}}^{-\frac{1}{2}}$, the optimization problem of Eq. (6) is transformed into finding the largest singular value.

Tensor Canonical Correlation Analysis (TCCA) [32] extends CCA to tensor-valued data to keep the original structure of the tensors. With

generality, suppose two n-dimensional tensors $\mathcal{X} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$ and $\mathcal{Y} \in \mathbb{R}^{d_1 \times \cdots \times d_m}$. TCCA seeks two tensors $\mathcal{A} = A_1 \circ \cdots \circ A_m$ of size $d_1 \times \cdots \times d_m$ and $\mathcal{B} = B_1 \circ \cdots \circ B_m$ of size $d_1 \times \cdots \times d_m$ that maximize the correlation between $\langle \mathcal{A}, \mathcal{X} \rangle$ and $\langle \mathcal{B}, \mathcal{Y} \rangle$, and Eq. (5) can be rewritten as Eq. (7)

$$\begin{aligned} &\arg\min_{\mathcal{A},\mathcal{B}} (\langle \mathcal{A}, \mathcal{X}_t \rangle - \langle \mathcal{B}, \mathcal{Y}_t \rangle)^2 \\ &\text{s.t.} \ \frac{1}{n} \sum_{t=1}^{n} \langle \mathcal{A}, \mathcal{X}_t \rangle^2 = \frac{1}{n} \sum_{t=1}^{n} \langle \mathcal{B}, \mathcal{Y}_t \rangle^2 = 1 \end{aligned} \tag{7}$$

where $\{\mathcal{A}_t, \mathcal{B}_t\}_{t=1}^{n}$ are the samples from $\mathcal{X}$, $\mathcal{Y}$. Eq. (7) can be solved by the idea of tensor decomposition. $\mathcal{X}$ can be projected into a low-dimensional space by one component A_h of $\mathcal{A}$, with the other components of $\mathcal{A}$ fixed. The partial contraction $\boldsymbol{X}_h$ of $\mathcal{X}$ with all components of $\mathcal{A}$ except A_h is defined as

$$\boldsymbol{X}_h = \mathcal{X} \times_2 A_1^T \cdots \times_h A_{h-1}^T \times_{h+2} A_{h+1}^T \cdots \times_{m+1} A_m^T \tag{8}$$

The partial contraction $\boldsymbol{Y}_h$ of $\mathcal{Y}$ is defined in the same way. $\mathcal{A}$ and $\mathcal{B}$ in Eq. (7) are regarded as existing in a low-rank space. There is a formula as follows

$$\begin{aligned} \boldsymbol{X}_{fh} = \sum_{t=1}^{n} (\mathcal{X}_t)_{(h+1)} \big(\boldsymbol{A}_{f-1,m} \cdots \otimes \boldsymbol{A}_{f-1,h+1} \otimes \\ \boldsymbol{A}_{f,h-1} \otimes \cdots \boldsymbol{A}_{f,1} \big) \\ \boldsymbol{Y}_{fh} = \sum_{t=1}^{n} (\mathcal{Y}_t)_{(h+1)} \big(\boldsymbol{B}_{f-1,m} \cdots \otimes \boldsymbol{B}_{f-1,h+1} \otimes \\ \boldsymbol{B}_{f,h-1} \otimes \cdots \boldsymbol{B}_{f,1} \big) \end{aligned} \tag{9}$$

where f is the current number of iterations, $\boldsymbol{A}_{f,h-1}$ represents $\boldsymbol{A}_{h-1}$ of the f-th iteration. Then, the following updating formula can be used to obtain $\mathcal{A}_f$ and $\mathcal{B}_f$.

$$\begin{aligned} \hat{\boldsymbol{A}_{fh}} &= (\boldsymbol{X}_{fh}^T \boldsymbol{X}_{fh} + c_x \boldsymbol{I})^{-1} \boldsymbol{X}_{fh}^T \boldsymbol{Y}_{fh} \boldsymbol{B}_{f-1,h}, \\ \boldsymbol{A}_{fh} &= \hat{\boldsymbol{A}_{fh}} \| \hat{\boldsymbol{A}_{fh}} \|^{-1} \\ \hat{\boldsymbol{B}_{fh}} &= (\boldsymbol{Y}_{fh}^T \boldsymbol{Y}_{fh} + c_y \boldsymbol{I})^{-1} \boldsymbol{Y}_{fh}^T \boldsymbol{X}_{fh} \boldsymbol{A}_{fh}, \\ \boldsymbol{B}_{fh} &= \hat{\boldsymbol{B}_{fh}} \| \hat{\boldsymbol{B}_{fh}} \|^{-1} \end{aligned} \tag{10}$$

where $c_x \boldsymbol{I}$ and $c_y \boldsymbol{I}$ are the regularization terms.

C. Trial Selection Tensor Canonical Correlation Analysis

Mutual information can measure the relationship between two random variables and provide a non-negative value. The higher the value, the higher the dependence between the two variables [71].

Pupil size variation is a measure of the brain's reactivity to emotional stimuli. In combination with facial expression, pupil dilation could potentially be an indicator of variation in emotional intensity [72] [73]. Previous studies always ignore the interaction between trials and provide redundant features. To avoid these drawbacks, TSTCCA measures the dependence between facial expression and pupil diameter trials by mutual information and selects a subset of trials with a strong correlation. The complementary information captured from facial expression and pupil diameter can eliminate trials unsuitable for depression recognition and optimize the original paradigm. After that, we conduct canonical correlation analysis between two modalities' feature tensors corresponding to the selected trial.

Suppose the two modalities feature tensors are inputs, $\mathcal{X} \in \mathbb{R}^{n \times m \times d_1}$ indicates facial expression features tensor, $\mathcal{Y} \in \mathbb{R}^{n \times m \times d_2}$ indicates pupil diameter features tensor. $\mathcal{X}$ and $\mathcal{Y}$ can be divided into n components $\mathcal{X} = \{\boldsymbol{X}_i\}_{i=1}^{n}$ and $\mathcal{Y} = \{\boldsymbol{Y}_i\}_{i=1}^{n}$, where

Algorithm 1: TSTCCA

Require: Facial expression tensor feature $\mathcal{X} \in \mathbb{R}^{n \times m \times d_1}$;
Pupil diameter tensor feature $\mathcal{Y} \in \mathbb{R}^{n \times m \times d_2}$;
Number of selected trials k;
Regularization parameters c_x and c_y.

1: **for** i=1,2...,n **do**
2: **for** j=1,2,...,m **do**
3: Calculate the mutual information $MI(\boldsymbol{X}_i, Y_{ij})$.
4: **end for**
5: Obtain the mutual information matrix of i-th subject $\boldsymbol{MI}_i$.
6: **end for**
7: Calculate the mean matrix for two modalities trials in Eq. (12).
8: Sort the values in MI_b in descending order.
9: Select the top k trials, obtain the selected facial expression features tensor $\tilde{\mathcal{X}}$ and the selected pupil diameter features tensor $\tilde{\mathcal{Y}}$.
10: Initialize R_x, R_y, L_x and L_y with random numbers.
11: Calculate $\boldsymbol{X}_R$, $\boldsymbol{Y}_R$, $\boldsymbol{X}_L$, and $\boldsymbol{Y}_L$ by Eq. (13).
12: **repeat**
13: Fix R_{xf} and R_{yf}, update L_{xf} and L_{yf}.
$\hat{L}_{xf} = (\boldsymbol{X}_{Rf}^T \boldsymbol{X}_{Rf} + c_x I)^{-1} \boldsymbol{X}_{Rf}^T \boldsymbol{Y}_{Rf} L_{y,f-1}$
$L_{xf} = \hat{L}_{xf} \| \hat{L}_{xf} \|^{-1}$
$\hat{L}_{yf} = (\boldsymbol{Y}_{Rf}^T \boldsymbol{Y}_{Rf} + c_y I)^{-1} \boldsymbol{Y}_{Rf}^T \boldsymbol{X}_{Rf} L_{xf}$
$L_{yf} = \hat{L}_{yf} \| \hat{L}_{yf} \|^{-1}$
14: Fix L_{xf} and L_{yf}, update R_{xf} and R_{yf}.
$\hat{R}_{xf} = (\boldsymbol{X}_{Lf}^T \boldsymbol{X}_{Lf} + c_x I)^{-1} \boldsymbol{X}_{Lf}^T \boldsymbol{Y}_{Lf} R_{y,f-1}$
$R_{xf} = \hat{R}_{xf} \| \hat{R}_{xf} \|^{-1}$
$\hat{R}_{yf} = (\boldsymbol{Y}_{Lf}^T \boldsymbol{Y}_{Lf} + c_y I)^{-1} \boldsymbol{Y}_{Lf}^T \boldsymbol{X}_{Lf} R_{xf}$
$R_{yf} = \hat{R}_{yf} \| \hat{R}_{yf} \|^{-1}$
15: $f = f + 1$
16: **until** (converged)
17: Project X and Y with X_R, Y_R, X_L, and Y_L.
$X' = L_x^T \tilde{\mathcal{X}} R_x, Y' = L_y^T \tilde{\mathcal{Y}} R_y$
18: Horizontally concatenate X' and Y'.
$W = [X'; Y']$

Ensure: The fusion feature matrix W.

$\boldsymbol{X}_i \in \mathbb{R}^{m \times d_1}$ and $\boldsymbol{Y}_i \in \mathbb{R}^{m \times d_2}$ represent the i-th subject of facial expression features tensor and pupil diameter features tensor, respectively. For each $\boldsymbol{X}_i$ and $\boldsymbol{Y}_i$, there are formula $\boldsymbol{X}_i = \{X_{ij}\}_{j=1}^{m}$ and $\boldsymbol{Y}_i = \{Y_{ij}\}_{j=1}^{m}$, where $X_{ij} \in \mathbb{R}^{d_1}$ and $Y_{ij} \in \mathbb{R}^{d_2}$ represent the j-th trial of the i-th subject of facial expression features tensor and the j-th trial of the i-th subject of pupil diameter features tensor, respectively. The formula of mutual information between facial expression features tensor and pupil diameter features tensor is

$$I(\mathcal{X}; \mathcal{Y}) = \sum_{\boldsymbol{X} \in \mathcal{X}} \sum_{\boldsymbol{Y} \in \mathcal{Y}} p(\boldsymbol{X}, \boldsymbol{Y}) lg \frac{p(\boldsymbol{X}, \boldsymbol{Y})}{p(\boldsymbol{X}) p(\boldsymbol{Y})} \tag{11}$$

where $p(\boldsymbol{X}, \boldsymbol{Y})$ is the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y})$, $p(\boldsymbol{X})$ and $p(\boldsymbol{Y})$ are the marginal distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

The process of the trial selection canonical correlation analysis (TSTCCA) method is shown in Algorithm 1. For each subject, we calculate the crossing mutual information $MI(\boldsymbol{X}_i, Y_{ij}) \in \mathbb{R}^{1 \times m}$ for each pupil diameter trial Y_{ij} and m facial expression trials $\boldsymbol{X}_i$ (in lines 2-4), which can explore the implied relationship between pupil and expression at different moments. Since there are m pupil diameter trials in total, the algorithm performs m mutual information calculations for each subject. Then, we concatenate m vectors $MI(\boldsymbol{X}_i, Y_{ij}), j = 1, 2, \cdots, m$ into a matrix $\boldsymbol{MI}_i \in \mathbb{R}^{m \times m}$ (in

line 5), which is the mutual information matrix of the i-th subject. For the i-th subject, the a-th column of $\boldsymbol{MI}_i$ represents the mutual information between the a-th trial of the facial expression and all trials of the pupil diameter, and b-th row represents the mutual information between the b-th trial of pupil diameter and all trials of the facial expression.

After that, we obtain n matrices $\boldsymbol{MI}_i, i = 1, 2, \cdots, n$ of n subjects, all of n $\boldsymbol{MI}_i$ stacked vertically and horizontally respectively to form the mutual information matrix $\boldsymbol{MI_{col}} \in \mathbb{R}^{(n\times m)\times m}$ and $\boldsymbol{MI_{row}} \in \mathbb{R}^{m\times(n\times m)}$, which are shown in Fig. 5. Vertical stacking and horizontal stacking are equivalent to joining together the columns and rows of all $\boldsymbol{MI}_i, i = 1, 2, \cdots, n$, respectively. The meaning of $\boldsymbol{MI}_{col}$ and $\boldsymbol{MI}_{row}$ correspond to the row and column interpretation of $\boldsymbol{MI}_i$, respectively, except that $\boldsymbol{MI}_i$ is based on a single subject, while $\boldsymbol{MI}_{col}$ and $\boldsymbol{MI}_{row}$ are based on all subjects, and are more general. For example, for n subjects, the a-th column of $\boldsymbol{MI}_{col}$ represents the mutual information between the a-th trial of the facial expression and all trials of the pupil diameter. Apparently, for all subjects, when a specific expression occurs in a trial, vertical stacking can find whether there is a relevant pupil change in other trials, while horizontal stacking, in contrast, can explore facial behavior at different moments associated with a particular pupil change.

The arithmetic mean can represent the average level of a set of data, so $\boldsymbol{MI_{col}}$ and $\boldsymbol{MI_{row}}$ are averaged by row and column, respectively, and the mean matrix $MI_{cm} \in \mathbb{R}^{1\times m}$ and $MI_{rm} \in \mathbb{R}^{1\times m}$ are obtained. The average of MI_{cm} and MI_{rm} is calculated as follows

$$MI_b = \frac{MI_{cm} + MI_{rm}}{2} \tag{12}$$

Moreover, we select top k trials with maximum mutual information. Two modalities feature tensors $\mathcal{X}$, $\mathcal{Y}$ after trial selection are named $\tilde{\mathcal{X}} \in \mathbb{R}^{n\times k\times d_1}$, $\tilde{\mathcal{Y}} \in \mathbb{R}^{n\times k\times d_2}$.

We seek right transforms $R_x \in \mathbb{R}^{d_1\times 1}$ and $R_y \in \mathbb{R}^{d_2\times 1}$, and left transforms $L_x \in \mathbb{R}^{k\times 1}$ and $L_y \in \mathbb{R}^{k\times 1}$ to maximize the correlations between $X' = L_x^T \tilde{\mathcal{X}} R_x$ and $Y' = L_y^T \tilde{\mathcal{Y}} R_y$. For $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ are three-dimensional tensors, $\boldsymbol{X}_{fh}$ and $\boldsymbol{Y}_{fh}$ in Eq. (9) can be simplified to the following formula

$$\begin{aligned} \boldsymbol{X}_R &= \tilde{\mathcal{X}} R_x \qquad & \boldsymbol{Y}_R &= \tilde{\mathcal{Y}} R_y \\ \boldsymbol{X}_L &= \tilde{\mathcal{X}}^{\dagger} L_x \qquad & \boldsymbol{Y}_L &= \tilde{\mathcal{Y}}^{\dagger} L_y \end{aligned} \tag{13}$$

where $\tilde{\mathcal{X}}^{\dagger}$ and $\tilde{\mathcal{Y}}^{\dagger}$ are the tensors of transposing the first and second dimensions of $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$, respectively. We use Eq. (10) to update R_x, R_y, L_x, and L_y iteratively until convergence.

After obtaining the left and right transforms, the original feature tensors $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ are projected into a new feature space. Finally, X' and Y' are concatenated along the horizontal to produce a multimodal feature matrix $W = [X'; Y']$, which is the input of classifiers.

V. Experiment and Results

In this section, we first introduce the preprocessing and feature extraction of the data. Then we propose three different trial selection comparison strategies (strategy 2, strategy 3, and strategy 4) to prove the reliability of strategy 1 in Algorithm 1 (in lines 1-8). Because the value of k (the number of selected trials) in Algorithm 1 is undetermined, we discuss the performance of different numbers of trials selected by four strategies and discover that strategy 1 of TSTCCA can provide the best performance with 25 trials. Finally, we detect depression by utilizing TSTCCA and compare this method with unimodal facial expression, unimodal pupil diameter, CCA fusion, and TCCA fusion to verify the superiority of our method.

We use kNN, SVM, and RF algorithms for the final classification, whose parameter ranges are shown in Table II. The hold-out method is applied 100 times for each experiment, and the accuracies are averaged to obtain a general result. At each time of classification, the 106 subjects in the data set are randomly split into the training and test sets, with sample sizes of 74 and 32, respectively. Furthermore, we use cross-validation to implement the hyper-parameter optimization involved in classifiers.

TABLE II: Parameter ranges of three classifiers.

Classifier	Parameter		Range
SVM	RBF kernel	gamma	$10^{-8} \sim 10^{8}$
		C	$10^{-8} \sim 10^{8}$
	linear kernel	C	$10^{-8} \sim 10^{8}$
	poly kernel	degree	$2 \sim 4$
kNN	$\mathrm{n}_{neighbors}$		$1 \sim 11$
RF	$\mathrm{n}_{estimators}$		$1 \sim 20$
	max_{depth}		$1 \sim 20$

A. Preprocess

The pupil diameter data are extracted from raw eye movement video by Pupil Player [74], which includes timestamps and confidence scores. Next, the pupil diameter data of each subject are divided into 40 segments according to the timestamps corresponding to all trials in the paradigm. The pupil diameter data with confidence scores less than 0.6 are removed to improve the credibility of the data. The pupil diameter data corresponding to the 20th, 38th, and 39th trials are removed because their overall confidence scores are too low. Therefore, 37 trials are retained in the final data.

Corresponding to the pupil diameter data, the facial expression data of each subject are also divided into 40 segments, and the 20th, 38th, and 39th segments of expression data are removed. The frame rate of the input facial expression video is 30 fps, and each trial lasts 5 seconds, thus yielding 150 frames per trial. Due to the inconsistency of frame rates, only 100 frames are taken for each trial. Finally, the Face-Alignment [75] face localization technique is used to crop the face region of each frame.

B. Feature Exaction

Multiple studies have proved that HOG is more suitable for describing facial expressions than other hand-crafted features [76] [77]. Thus, for facial expression data, we first extract a HOG [78] feature $X \in \mathbb{R}^{288\times 1}$ from each frame. The HOG feature parameters are set as follows: the window size is (32,64), the block size is (16,16), the block stride is (16,16), the cell size is (8,8), and the number of bins is 9. Since each trial contains 100 frames of images, each trial includes 100 HOG features. We calculate the average value of 100 HOG features on each trial to reduce the dimensionality of features. Finally, each trial is represented by only one HOG average feature, so HOG features extracted from 106 subjects are denoted by $\mathcal{X} \in \mathbb{R}^{106\times 37\times 288}$.

The eye tracker with a frame rate of 200 collects pupil images, and pupil-lab [74] does the follow-up analysis. Each frame of the pupil image corresponds to a pupil diameter value and finally forms a time series of pupil diameter. We refer to the features described in Table 2 of [79] for this time series. According to the characteristics of our data set, 8 time domain features (p_1-p_4, p_7, p_9, p_{11}, and peak-to-peak value) and 12 frequency domain (p_{12}-p_{23}) features are selected. Then, each trial is represented by a feature vector $Y \in$

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3322271

$\mathbb{R}^{20\times1}$ in the time and frequency domains. Finally, features in the time and frequency domains extracted from 106 subjects are denoted by $\mathcal{Y} \in \mathbb{R}^{106\times37\times20}$.

C. TSTCCA with Different Numbers of Trials

The number of selected trials (k) affects the performance of TSTCCA, so we need to seek the optimal k. The mutual information calculation method introduced in Algorithm 1 (in lines 1-8) is named trial selection strategy 1. To demonstrate its superiority, we devise three other trial selection strategies.

(1) Strategy 2: In line 3 of Algorithm 1, strategy 1 calculates crossing mutual information $MI(\boldsymbol{X}_i, Y_{ij}) \in \mathbb{R}^{1\times m}$ for each pupil diameter trial Y_{ij} and m facial expression trials $\boldsymbol{X}_i$. Unlike strategy 1, strategy 2 calculates one-to-one mutual information $MI(X_{ij}, Y_{ij}) \in \mathbb{R}^1$ for single pupil diameter trial Y_{ij} and single facial expression trial X_{ij}. Crossing mutual information introduces the connection between two modalities in different time periods into trial selection and is not limited to strict time synchronization. The comparison of the two types of mutual information is shown in Fig. 4. Similar to strategy 1, strategy 2 concatenates m $MI(X_{ij}, Y_{ij})$ into a vector $MI_i \in \mathbb{R}^{1\times m}$ in line 5. In line 7, it uses the vertically stacking method to obtain the mutual information matrix $\boldsymbol{MI_{all}} \in \mathbb{R}^{n\times m}$, and averages $\boldsymbol{MI_{all}}$ by row. Finally, it obtains the mean matrix $MI_u \in \mathbb{R}^{1\times m}$.

(2) Strategy 3: The first 6 lines of Algorithm 1 are the same as strategy 1, strategy 3 calculates crossing mutual information. In line 7, strategy 1 stacks n $\boldsymbol{MI}_i$ vertically and horizontally respectively to obtain $\boldsymbol{MI_{col}}$ and $\boldsymbol{MI_{row}}$, and averages them by row and column respectively to get MI_{cm} and MI_{rm}. Finally, the arithmetic mean of MI_{cm} and MI_{rm} is used as the basis for trial selection. Strategy 3 takes a simpler way, it only needs MI_{cm} to select trials. The g-th element of MI_{cm} indicates the mutual information between the features corresponding to all pupil diameter trials of all subjects and the features corresponding to the g-th facial expression trial.

(3) Strategy 4: Similar to strategy 3, it obtains the vector MI_{rm} in line 7. The g-th element of MI_{rm} indicates the mutual information between the features corresponding to all facial expression trials of all subjects and the features corresponding to the g-th pupil diameter trial. The differences between strategy 3 and strategy 4 are shown in Fig. 5.

To determine the value of k, we employ TCCA method as a basis. The k of TCCA equals 37, which means it does not make any trial selection. We utilize four trial selection strategies to get subsets of 37 original trials respectively, and focus on comparing the performance of strategy 1 with other strategies. Fig. 6 illustrates the highest classification accuracies of four strategies with different values of k. From Fig. 6, strategy 1 achieves the highest accuracy of 78.81%, when selecting 25 trials. The possible reason is that strategy 1 takes the arithmetic average of mutual information of strategy 3, and strategy 4, and inherits valuable information from both strategies. In addition, the classification accuracies of the four strategies are always higher than the baseline method TCCA. According to the above results, we select 25 trials for the subsequent analysis.

The 37 trials in the original paradigm contain 10 positive stimulus trials, 7 negative stimulus trials, and 20 neutral stimulus trials. The 25 selected trials in strategy 1 contain 8 positive stimulus trials, 5 negative stimulus trials, and 12 neutral stimulus trials, which account for 80%, 71.43%, and 60% of the total positive, neutral, and negative stimulus trials, respectively. Apparently, in the selected trials, there is a greater proportion of positive and negative stimuli than neutral stimuli. It has been proved that depressed patients pay significantly different attention to negative and positive information than normal controls. The attention bias to negative information of depressed patients has a relatively consistent result and is affected by the severity of the depression [80], [81]. Our paradigm of free viewing of emotional pictures is a sort of visual task. Regarding visual tasks, [82] used emotional face stimulation and found that depressed

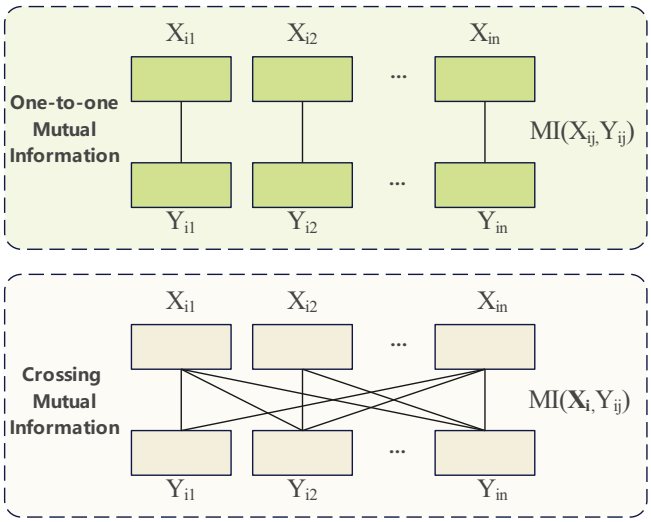


Fig. 4: Two option to calculate mutual information. The first figure is the one-to-one mutual information, which calculates the mutual information between one-to-one correspondence trials of two modalities. The second figure is the crossing mutual information, which crosses to calculate the mutual information between trials of two modalities.

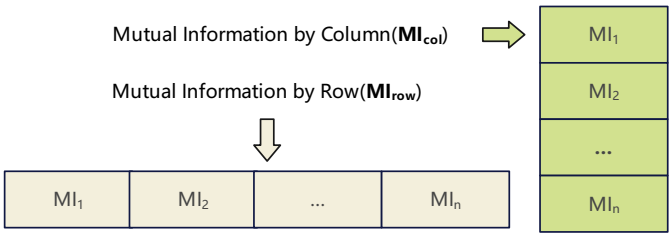


Fig. 5: Two options to stack mutual information matrices. $\boldsymbol{MI_{col}}$ stacks $MI_i, i = 1, 2, \cdots, n$ vertically, $\boldsymbol{MI_{row}}$ stacks $MI_i, i = 1, 2, \cdots, n$ horizontally.

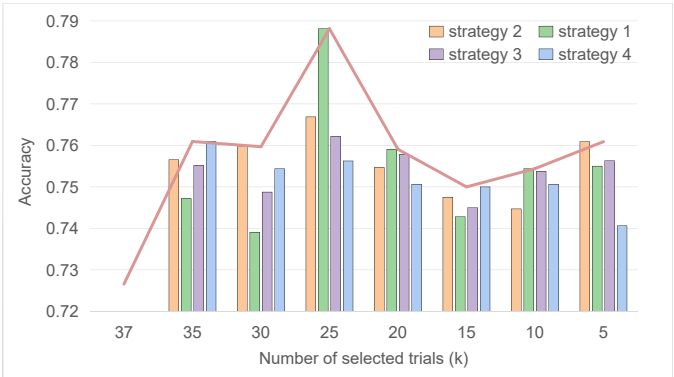


Fig. 6: Best classification performance of TSTCCA when selecting different numbers of trials. Each element in the histogram represents the accuracy corresponding to the best-performing classifier. Each point in the line represents the accuracy corresponding to the best-performing classifier and selection strategy.

patients had directed their attention selectively to negative faces. Some scholars have also studied the effect of positive stimulation on patients with depression. Patients with more severe depression demonstrated stronger negative mood prediction biases and weaker positive mood prediction biases [83]. [84] studied the biases between depressed patients and normal controls using both verbal and pictorial stimuli and found that depressed patients reduce perceptual sensitivity to positive pictures and words. Previous studies have verified the reliability of our choice of k=25.

D. Comparison of Unimodal and Other Multimodal Fusion Methods

To verify the superiority of TSTCCA, we compare the performance of unimodal facial expression, unimodal pupil diameter, CCA fusion, and TCCA fusion. For CCA fusion, there is one parameter $n_{component} = 2$, which is the number of components to keep. For TCCA fusion and TSTCCA fusion, there are three parameters $n_{iter} = 100$, $c_x = 0.01$, and $c_y = 0.01$, which are the maximum number of iterations and regularization parameters of x and y. Moreover, $k = 25$ in TSTCCA, which means TSTCCA selects 25 trials for analysis. SVM, kNN, and RF algorithms are used for classification. The box plots in Fig. 7 show the distribution of classification accuracies. The classification accuracies of the above methods obtained by SVM are generally the best. The accuracies of unimodal facial expression and unimodal pupil diameter are only 59.38% and 64.78%, respectively. Among the multimodal fusion method, TSTCCA has the best performance, whose accuracy is 78.81%. Relatively speaking, when 100 classifications are performed, the accuracy distribution of TSTCCA is more concentrated, which indicates that it can maintain a stable and excellent classification performance.

Confusion matrix [85] is an index to evaluate the classification results of the model, each column of the matrix represents an instance prediction of a class, while each row represents an instance of an actual class. Sensitivity (also known as recall rate) represents the proportion of the actual detected positive instances to the total positive instances, and specificity represents the proportion of the actual detected negative instances to the total negative instances. Sensitivity and specificity measure the capacity of the classifier to recognize positive and negative instances, respectively.

Fig. 8 is the confusion matrices of SVM. TSTCCA has the highest precision of 84.38%, which indicates that the fusion representation extracted by TSTCCA can significantly reflect the facial and pupillary changes of depressed subjects. The classification results of three classifiers are shown in Table III, Table IV, and Table V. Regardless of which classifier is used, TSTCCA consistently achieves the highest F1 values, and it has a specificity of 87.51% with SVM classifier. These facts prove our model can be used for preliminary screening of depression to a certain extent. In general, multimodal results are superior to unimodal results, and TSTCCA performs the best given its advantages. We use the paired sample t-test to assess the differences between TSTCCA and other methods and find significant differences, which reveals that TSTCCA has a significant improvement over other methods.

VI. Discussion

A. Comparative Analysis of Unimodal and Multimodal Fusion

When we only consider the unimodal features, the classification performance is always unsatisfactory. Each model has advantages and disadvantages. Just as facial expression directly reflects an individual actual emotional state that can hardly be disguised, and pupil diameter reflects the physiological response of the subject after the stimulus.

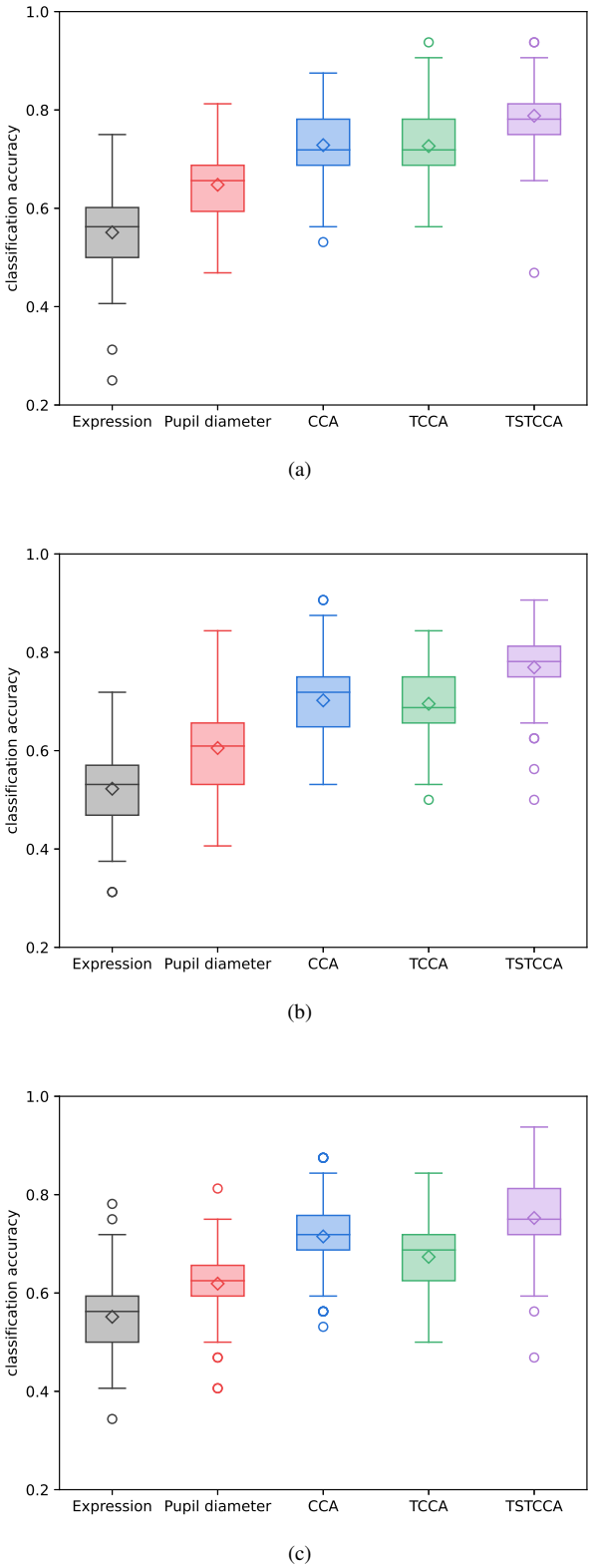
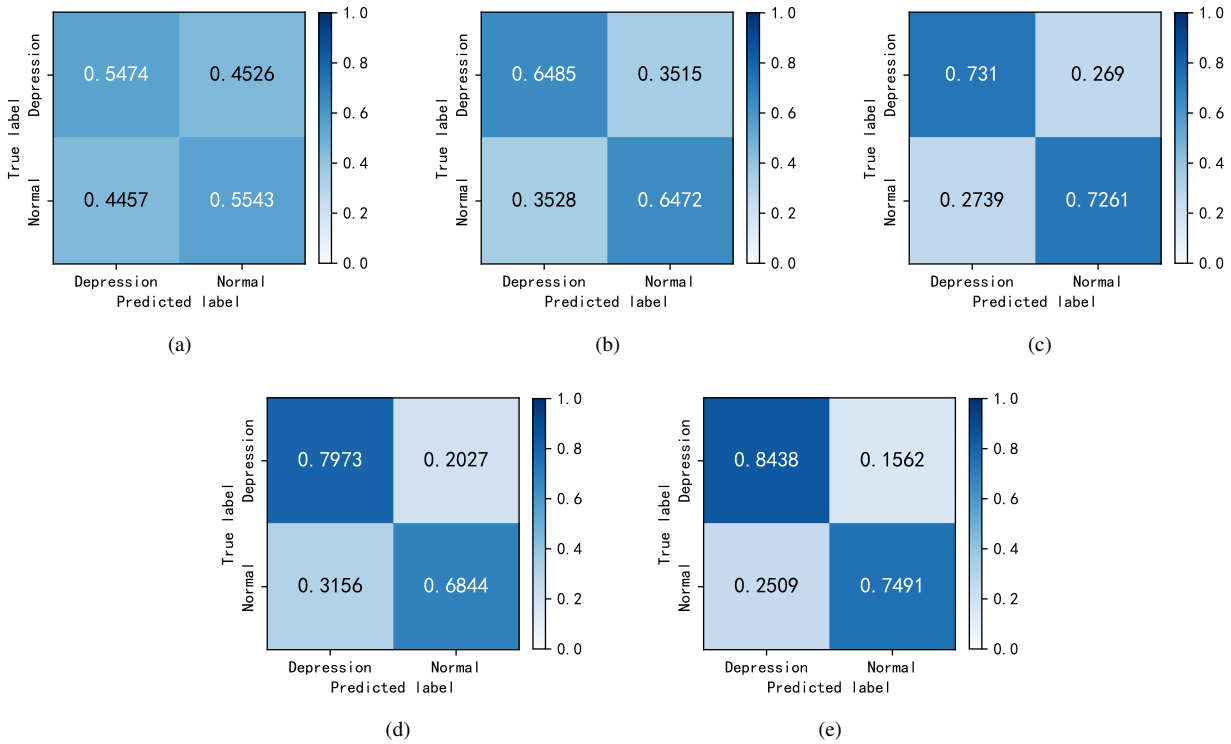


Fig. 7: Box plots of the classification accuracies obtained by unimodal facial expression, unimodal pupil diameter, CCA fusion, TCCA fusion, and TSTCCA fusion. (a) is the classification result of SVM, (b) is the classification result of kNN, (c) is the classification result of RF.

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2023.3322271



(a) (b) (c) (d) (e)

Fig. 8: Confusion matrices of (a) unimodal facial expression, (b) unimodal pupil diameter, (c) CCA fusion, (d) TCCA fusion, and (e) TSTCCA fusion by SVM.

TABLE III: Evaluation criteria (accuracy, sensitivity, specificity, and F1-score) of unimodal facial expression and pupil diameter, as well as multimodal CCA, TCCA, and TSTCCA using the SVM classifier and indications of significant difference (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score
Facial expression	59.38***	55.51**	57.44***	0.55***
Std.	7.60	15.95	13.00	0.09
Pupil diameter	64.78***	63.88***	67.16***	0.63***
Std.	8.81	17.24	17.90	0.10
CCA	72.84***	**71.69**	74.47***	0.72***
Std.	7.35	12.43	11.91	0.09
TCCA	72.66***	60.43***	85.25*	0.68***
Std.	6.89	11.56	11.28	0.09
TSTCCA	**78.81**	70.34	**87.51**	**0.76**
Std.	6.67	11.66	10.42	0.08

TABLE IV: Evaluation criteria (accuracy, sensitivity, specificity, and F1-score) of unimodal facial expression and pupil diameter, as well as multimodal CCA, TCCA, and TSTCCA using the kNN classifier and indications of significant difference (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score
Facial expression	52.25***	64.19*	41.97***	0.56***
Std.	8.53	13.86	14.90	0.08
Pupil diameter	60.53***	66.57**	55.73***	0.62***
Std.	8.05	13.07	14.70	0.09
CCA	70.22***	**72.03***	69.67***	0.70***
Std.	7.93	10.98	14.19	0.08
TCCA	69.53***	58.84***	81.11*	0.65***
Std.	6.76	13.65	12.87	0.09
TSTCCA	**76.94**	69.12	**84.87**	**0.74**
Std.	6.67	11.41	10.84	0.08

Based on the demand of clinical application, we must combine the beneficial information of multimodal to improve the recognition rate of depressed patients. The ability of a multimodal method to outperform a unimodal method is well established in the literature [86].

The experimental results verify this conclusion. Fig. 7 reveals that all multimodal methods fusing facial expression and pupil diameter have high accuracy, and TSTCCA performs the best. Fig. 8 also demonstrates that multimodal fusion methods have better precision than unimodal methods, and TSTCCA has the best performance. Moreover, Table III, Table IV, and Table V list various evaluation criteria to prove that TSTCCA is superior to other methods. TSTCCA solves the problem of feature redundancy by selecting trials with strong dependence between facial expression and pupil diameter. TSTCCA makes full use of the complementary information of multimodality and maximizes the cross-correlation between multimodal features.

B. Opportunities for Paradigm Optimization Observed in the Trial Subset

It has been proved that TSTCCA improves performance in depression recognition. Besides the classification results, we are concerned about the distribution of stimulus types in depression predictions based on a subset of selected trials for the purpose of paradigm

TABLE V: Evaluation criteria (accuracy, sensitivity, specificity, and F1-score) of unimodal facial expression and pupil diameter, as well as multimodal CCA, TCCA, and TSTCCA using the RF classifier and indications of significant difference (*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$).

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score
Facial expression	55.16***	56.11***	55.44***	0.54***
Std.	7.64	13.94	12.29	0.11
Pupil diameter	61.91***	59.12***	65.50***	0.60***
Std.	8.09	15.50	14.88	0.11
CCA	71.50***	70.50***	73.41	0.70***
Std.	7.54	12.71	13.15	0.09
TCCA	67.34***	59.09***	**76.60****	0.63***
Std.	6.65	13.83	14.50	0.09
TSTCCA	**75.25**	**78.61**	72.56	**0.75**
Std.	7.29	12.60	11.33	0.09

TABLE VI: The distributions of stimulus types in the trial subsets selected by the four strategies. The table records the number of positive, neutral, and negative trials in the optimal trial subset and their percentages of the total number of positive, neutral, and negative trials, respectively. k indicates the number of selected trials.

	strategy 1	strategy 2	strategy 3	strategy 4
k	25	25	25	35
Positive	8 (0.80)	7 (0.70)	8 (0.80)	10 (1.00)
Negative	5 (0.71)	6 (0.85)	5 (0.71)	7 (1.00)
Neutral	12 (0.60)	12 (0.60)	12 (0.60)	18 (0.90)

optimization. To improve the persuasiveness of the conclusion, we not only analyze strategy 1 in TSTCCA, but also discuss the other three strategies. Specifically, we discuss the significance of the accuracy trend in Fig. 6 in conjunction with the identifiers of trials obtained by four selection strategies. We will explore this question in terms of the highest accuracies of different strategies and the trend of the broken lines.

For strategies 1, 2, 3, and 4, the highest accuracy is 78.81%, 76.69%, 76.22%, and 76.09%. Notably, although the trial subsets selected by the four strategies are different, they share 18 topics. The identifiers of these 18 trials are 10, 8, 27, 40, 37, 16, 29, 3, 13, 22, 28, 25, 18, 17, 21, 12, and 23. It is reliable to design 25 trials and preferentially select these 18 trials for future paradigm optimization work. The distribution of stimulus types in the optimal trial subset obtained by four strategies is shown in Table VI. For strategies 1 and 3, the percentage of positive trials is the largest, followed by negative and neutral ones. In the case of strategy 2, the percentage of negative trials is greater than this of positive and neutral. In general, each optimal trial subset includes stimuli with positive, neutral, and negative attributes, which suggests that all three stimuli work in concert to more effectively discriminate between depressed and comparison individuals. Both positive and negative trials always have a greater probability of being selected in all optimal trial subsets, because the attentional biases of depressed patients differ from normal controls when facing both negative and positive stimuli.

By diving into the points in Fig. 6, we observe that several frequent identifiers in the deleted trials, as shown in Table VII. There are three crucial points of the line chart in Fig. 6, which are k = 35, k = 25, and k = 15. On the whole, when k decreases from 37 to 35, the accuracy is significantly improved, which is about 3% higher than TCCA. Then

TABLE VII: The frequent identifiers in the deleted trials obtained by the four selection strategies when k decreases. None * indicates two repetitions, * indicates three repetitions, ** indicates four repetitions.

	The frequent identifiers in the deleted trials
k:37→35	5, 11*
k:35→30	14, 19, 31*, 33, 36
k:30→25	7*, 9, 34
k:25→20	3, 12*, 15, 19, 23
k:20→15	17, 18, 21*, 25*
k:15→10	1, 16, 17, 22, 26, 28**
k:10→5	3, 6, 16, 27, 32*
k:5→0	4, 8*, 10*, 27, 29, 37, 40

the accuracy goes down at k = 30. Afterward, the accuracy reaches the peak at k=25. When k continues to decrease, the accuracy also decreases. Until k=15, the accuracy will slowly back up. Among the above processes, when k=35, 25, 10, the accuracy increases, thus the deleted trials have few benefits on depression recognition. When k=30, 20, 15, the accuracy decreases, so the deleted trials should be retained as appropriate.

These results suggest that the 25 trials with more negative and positive stimuli are crucial to support the diagnosis of depression, which may simplify the paradigm by preferentially removing trials in Table VII. Table VII shows the trial identifiers that are deleted by multiple strategies when the value of k is reduced. For example, the first row of the table shows that two strategies delete the fifth trial (identifier 5) and three strategies delete the eleventh trial (identifier 11) when k decreases from 37 to 35. However, it is hard to capture humans' complex psychological activities in a short experiment. But anyway, this study captures the key information related to depression, and the physiological and behavioral changes associated with these 25 trials should be paid extra attention.

VII. Conclusion

In the specific experiment of depression recognition, we propose TSTCCA to utilize the correlation and complementary relationship between facial expression and pupil diameter. Previous studies always ignore the spatiotemporal correlation between facial expression and pupil diameter, which might lead to high bias and high variance of classification models. TSTCCA fuses the multimodal high relevant features to optimize the feature space and improve the robustness of classification. Moreover, we find that negative and positive trials are associated with higher arousal for depressive facial behavioral phenotypes, which is in accordance with previous studies that depressed patients had attentional avoidance of positive stimuli and difficulty in attentional disengagement of negative stimuli. The preferred trial subset offers the possibility of paradigm optimization and is expected to support the clinical diagnosis of depression.

References

[1] Damian F Santomauro, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712, 2021.

[2] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R Merikangas, A John Rush, Ellen E Walters, and Philip S Wang. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *Jama*, 289(23):3095–3105, 2003.

[3] Zeinab Abd Elsalam Sarhan, Hanan Anwer El Shinnawy, Mohamed Elsayed Eltawil, Yassmin Elnawawy, Wegdan Rashad, and Mohammed Saadeldin Mohammed. Global functioning and suicide risk in patients with depression and comorbid borderline personality disorder. *Neurology, Psychiatry and Brain Research*, 31:37–42, 2019.

[4] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of Neurolinguistics*, 20(1):50–64, 2007.

[5] Junxin Chen, Shuang Sun, Li-bo Zhang, Benqiang Yang, and Wei Wang. Compressed sensing framework for heart sound acquisition in internet of medical things. *IEEE Transactions on Industrial Informatics*, 18(3):2000–2009, 2022.

[6] Gabrielle A Carlson and Frederick K Goodwin. The stages of mania: A longitudinal analysis of the manic episode. *Archives of general psychiatry*, 28(2):221–228, 1973.

[7] Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38, 2017.

[8] Zhijun Dai, Heng Zhou, Qingfang Ba, Yang Zhou, Lifeng Wang, and Guochen Li. Improving depression prediction using a novel feature selection algorithm coupled with context-aware analysis. *Journal of Affective Disorders*, 295:1040–1048, 2021.

[9] Aleks Stolicyn, J Douglas Steele, and Peggy Seriès. Prediction of depression symptoms in individual subjects with face and eye movement tracking. *Psychological medicine*, 52(9):1784–1792, 2022.

[10] Mariska E Kret and Annemie Ploeger. Emotion processing deficits: a liability spectrum providing insight into comorbidity of mental disorders. *Neuroscience & Biobehavioral Reviews*, 52:153–171, 2015.

[11] Alejandra Sel, Beatriz Calvo-Merino, Simone Tuettenberg, and Bettina Forster. When you smile, the world smiles at you: Erp evidence for self-expression effects on face processing. *Social cognitive and affective neuroscience*, 10(10):1316–1322, 2015.

[12] Helen Davies, I Wolz, J Leppanen, F Fernandez-Aranda, U Schmidt, and K Tchanturia. Facial expression to emotional stimuli in non-psychotic disorders: A systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 64:252–271, 2016.

[13] Michel Pierre Janisse. Pupil size, affect, and exposure frequency. *Social Behavior & Personality: an international journal*, 2(2), 1974.

[14] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[15] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

[16] Nitish Srivastava and Russ R Salakhutdinov. Multimodal learning with deep boltzmann machines. *Advances in neural information processing systems*, 25, 2012.

[17] Abrar H Abdulnabi, Bing Shuai, Zhen Zuo, Lap-Pui Chau, and Gang Wang. Multimodal recurrent neural networks with information transfer layers for indoor scene labeling. *IEEE Transactions on Multimedia*, 20(7):1656–1671, 2017.

[18] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921, 2019.

[19] Jing Zhu, Changlin Yang, Xiannian Xie, Shiqing Wei, Yizhou Li, Xiaowei Li, and Bin Hu. Mutual information based fusion model (mibfm): Mild depression recognition using eeg and pupil area signals. *IEEE Transactions on Affective Computing*, 2022.

[20] Tuan-Linh Nguyen, Swathi Kavuri, and Minho Lee. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. *Neural Networks*, 118:208–219, 2019.

[21] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[22] Pin Wang, En Fan, and Peng Wang. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognition Letters*, 141:61–67, 2021.

[23] Wei Wang, Xinhua Yu, Bo Fang, Dianna-Yue Zhao, Yongyong Chen, Wei Wei, and Junxin Chen. Cross-modality LGE-CMR segmentation using image-to-image translation based data augmentation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[24] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

[25] Zhonglin Lin, Changshui Zhang, Wei Wu, and Xiaorong Gao. Frequency recognition based on canonical correlation analysis for ssvep-based bcis. *IEEE transactions on biomedical engineering*, 53(12):2610–2614, 2006.

[26] Nicolle M Correa, Tulay Adali, Yi-Ou Li, and Vince D Calhoun. Canonical correlation analysis for data fusion and group inferences. *IEEE signal processing magazine*, 27(4):39–50, 2010.

[27] Xiaowei Zhang, Jing Pan, Jian Shen, Zia Ud Din, Junlei Li, Dawei Lu, Manxi Wu, and Bin Hu. Fusing of electroencephalogram and eye movement with group sparse canonical correlation analysis for anxiety detection. *IEEE Transactions on Affective Computing*, 2020.

[28] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[29] Ning Sun, Zhen-hai Ji, Cai-rong Zou, and Li Zhao. Two-dimensional canonical correlation analysis and its application in small sample size face recognition. *Neural Computing and Applications*, 19:377–382, 2010.

[30] Sun Ho Lee and Seungjin Choi. Two-dimensional canonical correlation analysis. *IEEE Signal Processing Letters*, 14(10):735–738, 2007.

[31] Tae-Kyun Kim, Shu-Fai Wong, and Roberto Cipolla. Tensor canonical correlation analysis for action classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[32] You-Lin Chen, Mladen Kolar, and Ruey S Tsay. Tensor canonical correlation analysis with convergence and statistical guarantees. *Journal of Computational and Graphical Statistics*, 30(3):728–744, 2021.

[33] Utkarsh Mahadeo Khaire and R Dhanalakshmi. Stability of feature selection algorithm: A review. *Journal of King Saud University-Computer and Information Sciences*, 2019.

[34] Peter Drotár, Juraj Gazda, and Zdenek Smékal. An experimental comparison of feature selection methods on two-class biomedical datasets. *Computers in biology and medicine*, 66:1–10, 2015.

[35] Stefan Scherer, Giota Stratou, Marwa Mahmoud, Jill Boberg, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. Automatic behavior descriptors for psychological disorder analysis. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[36] Heiner Ellgring. *Non-verbal communication in depression*. Cambridge University Press, 2007.

[37] Anastasia Pampouchidou, Panagiotis G Simos, Kostas Marias, Fabrice Meriaudeau, Fan Yang, Matthew Pediaditis, and Manolis Tsiknakis. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 10(4):445–470, 2017.

[38] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 11(3):542–552, 2018.

[39] Wheidima Carneiro De Melo, Eric Granger, and Abdenour Hadid. Depression detection based on deep distribution learning. In *2019 IEEE international conference on image processing (ICIP)*, pages 4544–4548. IEEE, 2019.

[40] Lang He, Mingyue Niu, Prayag Tiwari, Pekka Marttinen, Rui Su, Jiewei Jiang, Chenguang Guo, Hongyu Wang, Songtao Ding, Zhongmin Wang, et al. Deep learning for depression recognition with audiovisual cues: A review. *Information Fusion*, 80:56–86, 2022.

[41] Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar. Spectral representation of behaviour primitives for depression analysis. *IEEE Transactions on Affective Computing*, 13(2):829–844, 2020.

[42] Lang He, Jonathan Cheung-Wai Chan, and Zhongmin Wang. Automatic depression recognition using cnn with attention mechanism from videos. *Neurocomputing*, 422:165–175, 2021.

[43] Lang He, Chenguang Guo, Prayag Tiwari, Hari Mohan Pandey, and Wei Dang. Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence. *International Journal of Intelligent Systems*, 37(12):10140–10156, 2022.

[44] Wheidima Carneiro de Melo, Eric Granger, and Miguel Bordallo Lopez. Mdn: A deep maximization-differentiation network for spatio-temporal depression detection. *IEEE Transactions on Affective Computing*, 2021.

[45] Minqiang Yang, Yu Ma, Zhenyu Liu, Hanshu Cai, Xiping Hu, and Bin Hu. Undisturbed mental state assessment in the 5g era: a case study of depression detection based on facial expressions. *IEEE Wireless Communications*, 28(3):46–53, 2021.

[46] Thomas Suslow, Anja Husslack, Anette Kersting, and Charlott Maria Bodenschatz. Attentional biases to emotional information in clinical depression: A systematic and meta-analytic review of eye tracking findings. *Journal of Affective Disorders*, 274:632–642, 2020.

[47] Jing Zhu, Zihan Wang, Tao Gong, Shuai Zeng, Xiaowei Li, Bin Hu, Jianxiu Li, Shuting Sun, and Lan Zhang. An improved classification model for depression detection using eeg and eye tracking data. *IEEE transactions on nanobioscience*, 19(3):527–537, 2020.

[48] Dan Zhang, Xu Liu, Lihua Xu, Yu Li, Yangyang Xu, Mengqing Xia, Zhenying Qian, Yingying Tang, Zhi Liu, Tao Chen, et al. Effective differentiation between depressed patients and controls using discriminative eye movement features. *Journal of Affective Disorders*, 307:237–243, 2022.

[49] Ruizhe Shen, Qi Zhan, Yu Wang, and Huimin Ma. Depression detection by analysing eye movements on emotional images. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7973–7977. IEEE, 2021.

[50] Jitao Zhong, Dixin Wang, Hongtong Wu, Peng Wang, Minqiang Yang, Hong Peng, and Bin Hu. Filterable sample consensus based on angle variance for pupil segmentation. *Digital Signal Processing*, 130:103695, 2022.

[51] Li-Ming Zhao, Rui Li, Wei-Long Zheng, and Bao-Liang Lu. Classification of five emotions from eeg and eye movement signals: complementary representation properties. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 611–614. IEEE, 2019.

[52] Greg J Siegle, Stuart R Steinhauer, Edward S Friedman, Wesley S Thompson, and Michael E Thase. Remission prognosis for cognitive therapy for recurrent depression using the pupil: utility and neural correlates. *Biological psychiatry*, 69(8):726–733, 2011.

[53] Greg J Siegle, Eric Granholm, Rick E Ingram, and Georg E Matt. Pupillary and reaction time measures of sustained processing of negative information in depression. *Biological psychiatry*, 49(7):624–636, 2001.

[54] Neil P Jones, Greg J Siegle, and Darcy Mandell. Motivational and emotional influences on cognitive control in depression: A pupillometry study. *Cognitive, Affective, & Behavioral Neuroscience*, 15:263–275, 2015.

[55] Jikun Wang, Yaodong Fan, Xudong Zhao, and Nanhui Chen. Pupillometry in chinese female patients with depression: a pilot study. *International journal of environmental research and public health*, 11(2):2236–2243, 2014.

[56] Rebecca B Price, Dana Rosen, Greg J Siegle, Cecile D Ladouceur, Kevin Tang, Kristy Benoit Allen, Neal D Ryan, Ronald E Dahl, Erika E Forbes, and Jennifer S Silk. From anxious youth to depressed adolescents: Prospective prediction of 2-year depression symptoms via attentional bias measures. *Journal of abnormal psychology*, 125(2):267, 2016.

[57] Mingyue Niu, Jianhua Tao, Bin Liu, Jian Huang, and Zheng Lian. Multimodal spatiotemporal representation for automatic depression level detection. *IEEE Transactions on Affective Computing*, 2020.

[58] Wanqing Xie, Chen Wang, Zhixiong Lin, Xudong Luo, Wenqian Chen, Manzhu Xu, Lizhong Liang, Xiaofeng Liu, Yanzhong Wang, Hui Luo, et al. Multimodal fusion diagnosis of depression and anxiety based on cnn-lstm model. *Computerized Medical Imaging and Graphics*, 102:102128, 2022.

[59] Jing Zhu, Shiqing Wei, Xiannian Xie, Changlin Yang, Yizhou Li, Xiaowei Li, and Bin Hu. Content-based multiple evidence fusion on eeg and eye movements for mild depression recognition. *Computer Methods and Programs in Biomedicine*, 226:107100, 2022.

[60] Xinghao Yang, Weifeng Liu, Wei Liu, and Dacheng Tao. A survey on canonical correlation analysis. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2349–2368, 2019.

[61] Yuri Levin-Schwartz, Yang Song, Peter J Schreier, Vince D Calhoun, and Tülay Adalı. Sample-poor estimation of order and common signal subspace with application to fusion of medical imaging data. *NeuroImage*, 134:486–493, 2016.

[62] Zhengshi Yang, Xiaowei Zhuang, Christopher Bird, Karthik Sreenivasan, Virendra Mishra, Sarah Banks, Dietmar Cordes, and Alzheimer's Disease Neuroimaging Initiative. Performing sparse regularization and dimension reduction simultaneously in multimodal data fusion. *Frontiers in neuroscience*, 13:642, 2019.

[63] Baiying Lei, Siping Chen, Dong Ni, and Tianfu Wang. Discriminative learning for alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. *Frontiers in aging neuroscience*, 8:77, 2016.

[64] Nandakishor Desai, Abd-Krim Seghouane, and Marimuthu Palaniswami. Multisubject fmri data analysis via two dimensional multi-set canonical correlation analysis. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 468–471. IEEE, 2017.

[65] Songze Tang, Liang Xiao, Wei Huang, Pengfei Liu, and Huicong Wu. Pan-sharpening using 2d cca. *Remote sensing letters*, 6(5):341–350, 2015.

[66] Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Multimodal emotion recognition using deep canonical correlation analysis. *arXiv preprint arXiv:1908.05349*, 2019.

[67] Peter Lang and Margaret M Bradley. The international affective picture system (iaps) in the study of emotion and attention. *Handbook of emotion elicitation and assessment*, 29:70–73, 2007.

[68] Shuang Liu, Di Zhang, Jingjing Tong, Feng He, Hongzhi Qi, Lixin Zhang, and Dong Ming. Eeg-based emotion estimation using adaptive tracking of discriminative frequency components. In *2017 39th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2231–2234. IEEE, 2017.

[69] Qingxiang Wang, Huanxin Yang, and Yanhong Yu. Facial expression video analysis for depression detection in chinese patients. *Journal of Visual Communication and Image Representation*, 57:228–233, 2018.

[70] W. Guo, H. Yang, Z. Liu, Y. Xu, and B. Hu. Deep neural networks for depression recognition based on 2d and 3d facial expressions under emotional stimulus tasks. *Frontiers in Neuroscience*, 15:609760, 2021.

[71] Mohammed A Ambusaidi, Xiangjian He, Priyadarsi Nanda, and Zhiyuan Tan. Building an intrusion detection system using a filter-based feature selection algorithm. *IEEE transactions on computers*, 65(10):2986–2998, 2016.

[72] Christine Fawcett, Elisabeth Nordenswan, Santeri Yrttiaho, Tuomo Häikiö, Riikka Korja, Linnea Karlsson, Hasse Karlsson, and Eeva-Leena Kataja. Individual differences in pupil dilation to others' emotional and neutral eyes with varying pupil sizes. *Cognition and Emotion*, pages 1–15, 2022.

[73] Mariska E Kret, Agneta H Fischer, and Carsten KW De Dreu. Pupil mimicry correlates with trust in in-group partners with dilating pupils. *Psychological science*, 26(9):1401–1410, 2015.

[74] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pages 1151–1160, 2014.

[75] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[76] Pierluigi Carcagnì, Marco Del Coco, Marco Leo, and Cosimo Distante. Facial expression recognition and histograms of oriented gradients: a comprehensive study. *SpringerPlus*, 4(1):1–25, 2015.

[77] Junkai Chen, Zenghai Chen, Zheru Chi, Hong Fu, et al. Facial expression recognition based on facial components detection and hog features. In *International workshops on electrical and computer engineering subfields*, pages 884–888, 2014.

[78] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[79] Yaguo Lei, Zhengjia He, Yanyang Zi, and Qiao Hu. Fault diagnosis of rotating machinery based on multiple anfis combination with gas. *Mechanical systems and signal processing*, 21(5):2280–2294, 2007.

[80] Josh M Cisler, Kate B Wolitzky-Taylor, Thomas G Adams Jr, Kimberly A Babson, Christal L Badour, and Jeffrey L Willems. The emotional stroop task and posttraumatic stress disorder: a meta-analysis. *Clinical psychology review*, 31(5):817–828, 2011.

[81] Amanda M Epp, Keith S Dobson, David JA Dozois, and Paul A Frewen. A systematic meta-analysis of the stroop task in depression. *Clinical psychology review*, 32(4):316–328, 2012.

[82] Ian H Gotlib, Elena Krasnoperova, Dana Neubauer Yue, and Jutta Joormann. Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of abnormal psychology*, 113(1):127, 2004.

[83] Susan J Wenze, Kathleen C Gunthert, and Ramaris E German. Biases in affective forecasting and recall in individuals with depression and anxiety symptoms. *Personality and Social Psychology Bulletin*, 38(7):895–906, 2012.

[84] Ruth Ann Atchley, Stephen S Ilardi, Keith M Young, Natalie N Stroupe, Aminda J O'Hare, Steven L Bistricky, Elizabeth Collison, Linzi Gibson, Jonathan Schuster, and Rebecca J Lepping. Depression reduces perceptual sensitivity for positive words and pictures. *Cognition & emotion*, 26(8):1359–1370, 2012.

[85] James T Townsend. Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, 9(1):40–50, 1971.

[86] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.