Multimodal Deep Learning for Remote Stress Estimation Using CCT-LSTM

Sayyedjavad Ziaratnia Tipporn Laohakangvalvit Midori Sugaya
Peeraya Sripian
College of Engineering, Shibaura Institute of Technology
Tokyo, Japan
{am20008,tipporn,doly,peeraya}@shibaura-it.ac.jp

Abstract

Stress estimation is key to the early detection and mitigation of health problems, enhancing driving safety through driver stress monitoring, and improving human–robot interaction efficiency by adapting to user's stress levels. In this paper, we present a novel method for video-based remote stress estimation and categorization, which involves two separate experiments: one for stress task classification and another for multilevel stress classification. The method combines two deep learning approaches, the Compact Convolutional Transformer (CCT) and Long Short-Term Memory (LSTM), to form a CCT-LSTM pipeline. For each modality (facial expression and rPPG), a CCT model is used to extract features, followed by an LSTM block for temporal pattern recognition. In stress task classification, T1, T2, and T3 tasks from the UBFC-Phys dataset are used, utilizing sevenfold cross-validation. The results indicated a mean accuracy of 83.2% and an F1 score of 83.4%. For multilevel stress classification, the control (lower stress) and test (higher stress) groups from the same dataset were used with fivefold cross-validation, achieving a mean accuracy of 80.5% and an F1 score of 80.3%. The results suggest that our proposed model surpasses existing stress estimation methods by effectively using multimodal deep learning and the CCT-LSTM pipeline for precise, non-invasive stress detection and categorization, with applications in health monitoring, safety, and interactive technologies.

1. Introduction

Stress, defined as physical, emotional, or mental tension caused by significant environmental changes, has become an integral part of modern living [18]. It can lead to chronic physical and mental health issues, such as depression, anxiety, chronic fatigue syndrome, diabetes, and cardiovascular disease [1]. As reported by the American Psychological Association, half of American adults have been negatively affected by stress, indicating its pervasive influence [11]. Given the potential harm that stress can cause, early detection and monitoring of stress are important in preventing short-term health issues from developing into long-term conditions [10, 16]. Stress detection is crucial not only in healthcare, but also in various fields such as monitoring the state of the driver and human–robot interaction. In driver state monitoring, the impact of cognitive stress on driving performance and traffic safety can be evaluated and mitigated [2]. Similarly, in robotic systems, incorporating automatic cognitive stress assessment in the feedback loop can improve the system's usability and efficiency by adapting the robot's behavior to the user's cognitive state [25].

Three distinct methodologies exist in the field of stress detection: questionnaires, behavioral analysis, and physiological analysis. Traditionally, stress detection has relied on subjective questionnaires and clinical interviews [17]. However, these conventional methods have limitations in feasibility for continuous monitoring [17]. In stress estimation, behavioral analysis is based on non-invasive indicators such as facial expressions, head motion, and eye gaze [24]. Still, the reliability is compromised, because individuals can control these behaviors [29]. Conversely, physiological signal analysis aims to capture stress markers through a multitude of methods, including electrocardiography (ECG), photoplethysmography (PPG), blood volume pressure (BVP), electromyography (EMG), electrodermal activity (EDA), respiratory measurements (RSP), and skin temperature (SKT) evaluations [14]. Although these physiological indicators can provide an in-depth understanding of an individual's physiological response to stress, they also require direct physical contact, and often a specialized person to install the sensor.

Over the last decade, significant advancements in remote photoplethysmography (rPPG) [5, 7, 19, 26] have paved the way for a noninvasive method of stress estimation. Sabour *et al*. [22] made a significant contribution in this field by creating the UBFC-Phys dataset following the Trier Social Stress Test protocol and recording facial video, PPG, and EDA signals. Their study involved three tasks: rest (T1),

speech task (T2), and arithmetic task (T3). Additionally, two groups of participants were subjected to these tasks, which varied in difficulty and induced varying stress levels. The control (ctrl) group faced easier tasks, whereas the test group tackled more challenging tasks; the analysis revealed that the latter experienced higher stress levels due to increased task complexity. Sabour *et al*. applied machine learning methods for both stress task (T1/T2/T3) and level (ctrl and test groups) classification. In the case of stress task classification, they achieved a satisfactory accuracy of 85.48% for binary classification tasks, but a notable decline in performance, down to 63.09%, was observed when the task was expanded to a three-class (T1 vs. T2 vs. T3) task classification. In the context of stress level classification, Sabour et al. focused solely on distinguishing between the control and test groups, achieving an accuracy of 69.73%. However, no studies exploring multilevel stress classification, such as differentiating among the T1 (relax), control (lower stress), and test (higher stress) groups, have yet been conducted. Further research has mainly concentrated on stress task classification. A recent study by Zhang *et al*. [30], used contact-based PPG and EDA methods for stress task classification, achieving a maximum accuracy of 81.8% in binary classification but experiencing a significant decline to 55.8% in the three-class task classification (T1 vs. T2 vs. T3). These outcomes highlight the limitations of current approaches in accurately classifying stress tasks and levels when more than two categories are involved.

To overcome existing limitations, we introduced a novel, comprehensive multimodal deep learning approach for remote stress estimation using CCT-LSTM. The essence and novelty of the proposed method are principally threefold:

1. **Integration of techniques**: The method integrates two advanced deep learning techniques, Compact Convolutional Transformer (CCT) and Long Short-Term Memory (LSTM), into a single CCT-LSTM pipeline for remote stress estimation. This unique integration combines the benefits of both techniques: CCT for efficient feature extraction and LSTM for temporal pattern recognition, thus creating a more powerful and accurate pipeline for stress estimation.

2. **Bidirectional use of CCT-LSTM for multimodality:** One novel aspect is the bidirectional use of the CCT-LSTM framework to construct a multimodal system. The framework uses CCT-LSTM in two directions: one for processing facial expressions based on 478 landmarks and another for handling rPPG signals from video data. This two-directional approach improves the multimodal nature of the system, making it even more robust and accurate in remote stress estimation.

3. **Application to remote stress tasks**: The CCT-LSTM framework and preprocessing techniques are applied to two specific tasks: remote stress task classification and multilevel stress classification. This is a novel application of these techniques and represents a significant advancement in the field of remote stress monitoring. The ability to determine not only whether someone is stressed but also the level of stress they are experiencing is particularly novel and could have wide-ranging applications in telemedicine, remote working, and other fields.

2. Background

This section explains multilevel stress estimation using remote and contact-based PPG and the essential concepts in the research. Then, it discusses the role and significance of Transformers in the field of computer vision. This includes a comprehensive analysis of their inherent strengths and limitations, along with a review of ongoing research aiming to enhance their functionality. The final part of this section describes the method used to convert time-series data into images, which is a crucial step in analyzing time-series data through computer vision.

2.1. Stress Estimation Using PPG/rPPG

Research in the area of stress estimation using rPPG and heart rate variability (HRV) analysis has led to the development of groundbreaking noninvasive techniques for stress tracking. Bousefsaf *et al*. [3] proposed a new model that uses facial videos of individuals to obtain rPPG signals and HRV characteristics. They observed a close correlation between signals such as the third-order derivative of HRV, high-frequency (HF), and heart rate, once they were smoothed and compared with EDA. Their study, which involved 12 subjects, reported variances in the values obtained in calm and stressful periods, supporting the idea that BVP and EDA signals are associated with stress levels. Furthering this idea, Mitsuhashi *et al*. [20] suggested a means to assess four tiers of stress based on pulse rate variability features, extracted from facial videos using the rPPG method. They broke a new ground in using rPPG for measuring multiple stress states and used a K-nearest neighbors (KNN) model based on HRV indices such as Average NN intervals (AVNN), the root mean square of successive differences (RMSSD), the proportion of NN50 divided by the total NNs (pNN50), and normalized low-frequency over high-frequency ratio (nLF/HF). However, the obtained results indicated that the method could not achieve high accuracy, except for the relaxed state.

Sabour *et al*. [22] significantly contributed to remote stress estimation by creating the first public dataset for multilevel social stress that includes facial video and biosignals: the UBFC-Phys dataset. However, as described in the previ-

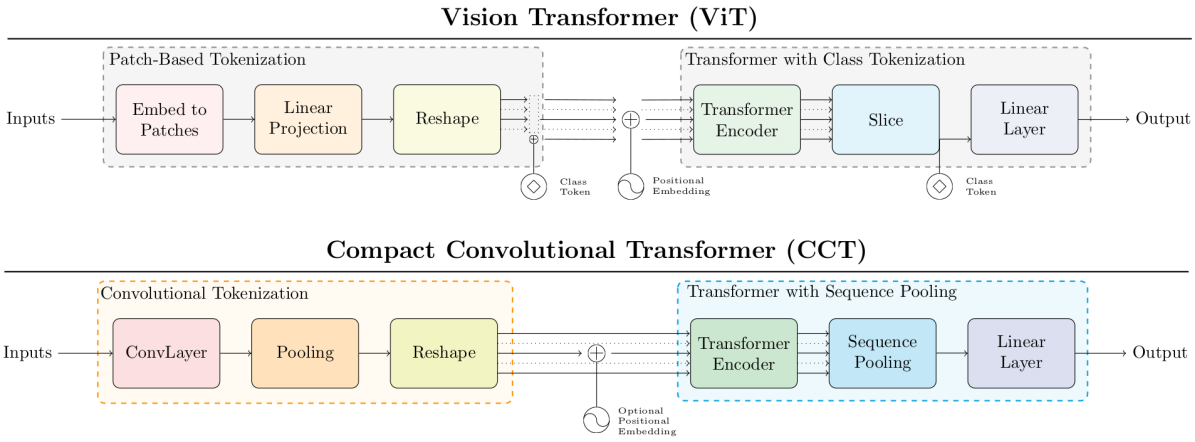

Figure 1. Pipeline comparison of Vision Transformer (ViT) and Compact Convolutional Transformer (CCT) [15]

ous section, stress task and level classifications still remain as challenging tasks.

While no recent attempts have been made in the multi-level stress classification on the UBFC-Phys dataset, Zhang *et al*. [30] conducted a study on stress task classification and introduced a multimodal stress detection framework based on a bidirectional cross and self-modal attention mechanism. This model was designed to integrate two physiological signals, BVP and EDA, while examining the temporal relationship between them. To evaluate the efficacy of their model in recognizing stress, Zhang *et al*. conducted comparative analyses with well-known neural networks specialized in stress detection using multimodal physiological signals, as well as several multimodal fusion models that use attention mechanisms. These comparisons were based on the UBFC-Phys dataset to categorize tasks into three levels: (1) T1 vs. T2, (2) T1 vs. T3, and (3) T1 vs. T2 vs. T3. The results of the sevenfold cross-validation indicated that Zhang *et al*. 's method achieved a mean accuracy of 81.8% for T1 vs. T2, 73.3% for T1 vs. T3, and 55.8% for T1 vs. T2 vs. T3.

2.2. Transformers in Computer Vision

Transformers, initially developed for natural language processing (NLP) tasks [6], have recently been applied successfully to computer vision problems. The Vision Transformer (ViT) [9], an adaptation of the original Transformer model, treats an image as a sequence of pixels or patches, similar to how a sentence is viewed as a sequence of words in NLP. The strength of the Transformers in computer vision lies in its global self-attention mechanism. This mechanism allows each part of an image (patch) to interact with all other parts, thereby capturing complex, long-range spatial dependencies within the image. Despite its promising results, Transformers have been criticized for their large computational and data requirements, often requiring training on large-scale datasets [15]. Furthermore, their efficacy with respect to temporal data is limited [23]. As such, while they traditionally excel in identifying spatial relationships, they often struggle to consistently recognize and analyze patterns over time. Transformers, in their native configuration, lack a built-in mechanism for temporal dependency, causing them to underperform in tasks involving sequential or time-series data. This constraint is due to the nonrecurrent nature of Transformers, making it difficult to maintain and understand the temporal continuity and causality inherently present in time-series data. The lack of a native temporal dimension understanding in the Transformer architecture is indeed a critical limitation when applied to problems where time-dependent correlations are important.

2.3. Compact Convolutional Transformer

The CCT [15] was developed to address the challenges of high computational cost and requirement of massive data that come with the ViT. The CCT offers a unique solution by integrating the global receptive field of Transformers and the local receptive field of convolutional neural networks (CNNs). As illustrated in Figure 1, the CCT bypasses the need for class tokens and positional embedding by implementing an innovative sequence pooling approach and utilizing convolutions. The model uses convolutional tokenization to process images as token sequences, which allows for efficient performance even with smaller datasets. The CCT applies an inductive bias through convolutional layers to minimize its dependence on positional embedding. The flexibility of the model allows it to adjust its size, capa-

ble of functioning with as few as 0.28M parameters, while still delivering high-quality results. Impressively, Hassani *et al*. [15] has demonstrated that the CCT outperforms the ViT in terms of accuracy on widely recognized computer vision benchmarks such as CIFAR-10 and ImageNet, all the while maintaining a lower computational cost.

2.4. Time-Series Data to Image Conversion

The Markov transition field (MTF) [27] is a method for representing time series data as an image-like structure, making it possible to analyze it using computer vision pioneers such as ViT or CCT. MTF works by calculating the transition probabilities between different states in the time series and organizing them in a 2D grid, resulting in an image that represents the transitions. Each pixel in the image represents the transition probability from one state to another. In this way, MTF captures both the temporal dependencies and value distribution characteristics of a time series. For a time-series such as $X = \{x_1, x_2, \cdots, x_n\}$, the values can be quantized in Q bins, and each x_i can be allocated to a related $q_j (j \in [1, Q])$. By calculating the transitions among bins in the way of a first-order Markov chain along each time step, a matrix W of $Q \times Q$ size is obtained. Its expression is as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1Q} \\ w_{21} & w_{22} & \cdots & w_{2Q} \\ \vdots & \vdots & \ddots & \vdots \\ w_{Q1} & w_{\mathrm{Q2}} & \cdots & w_{\mathrm{QQ}} \end{bmatrix} \tag{1}$$

$$w_{ij} = p\{x_t \in q_i \mid x_{t-1} \in q_j\} \tag{2}$$

where w_{ij} denotes the probability that an element in q_j is followed by an element in q_i. After normalization by $\sum_{j=1}^{Q} w_{ij} = 1$, W is considered to be the Markov transition matrix. Because W is not sensitive to the distribution of X and time dependency, in order to reduce the loss of information, the M_{ij} in the MTF is defined as follows:

$$\mathrm{M}=\begin{bmatrix} w_{ij} \mid x_1 \in q_i, x_1 \in q_j & \cdots & w_{ij} \mid x_1 \in q_i, x_n \in q_j \\ w_{ij} \mid x_2 \in q_i, x_1 \in q_j & \cdots & w_{ij} \mid x_2 \in q_i, x_n \in q_j \\ \vdots & \ddots & \vdots \\ w_{ij} \mid x_n \in q_i, x_1 \in q_j & \cdots & w_{ij} \mid x_n \in q_i, x_n \in q_j \end{bmatrix} \tag{3}$$

3. Our Proposed Framework

This study proposes a novel, multimodal deep learning framework specifically designed for remote stress estimation. We use a combination of facial expressions, using 478 distinctive landmarks, and rPPG signal extracted from video data based on the UBFC-Phys dataset [22]. As presented in Figure 2, the operational framework of our method is organized into a CCT-LSTM pipeline which combines the power of two profound deep learning methods: the CCT and LSTM.

3.1. Feature Extraction Using Compact Convolutional Transformer

The initial phase of our methodology involves the use of CCT. CCT is a deep learning method selected for its efficiency in extracting significant features from the datasets. As presented in Figure 2, this method serves a dual purpose in this study, aiding in the extraction of relevant features from both the 478 facial landmarks and the rPPG signal. Owing to its convolutional design, the CCT is inherently capable of effectively handling spatial information, which is particularly crucial when processing facial landmarks. Furthermore, the transformer component of the CCT aids in managing the intricate relationships between various features.

3.2. Temporal Pattern Recognition Using Long Short-Term Memory

Upon the completion of feature extraction, our method progresses to the phase of temporal pattern recognition, facilitated by LSTM. LSTM forms a class of recurrent neural networks, which have demonstrated superior capability in identifying patterns over time-series data. Given the temporal nature of stress markers and rPPG signal, the application of LSTM in stress estimation is particularly relevant. The LSTM aids in accurately capturing and interpreting the temporal patterns presented by the signal.

3.3. CCT-LSTM Pipeline

The CCT and LSTM form a unified pipeline, central to our methodology, which harnesses the benefits of both methods. As presented in Figure 2, the pipeline begins with the CCT processing of facial landmarks and rPPG signal to extract relevant features. Following this, the LSTM takes over to analyze the temporal patterns embedded within these features. This step-by-step progression ensures a comprehensive analysis of the two modalities, resulting in a more accurate and reliable stress estimation.

4. Method

4.1. Dataset

The UBFC-Phys dataset [22] was specifically created for psycho-physiological research and was collected from 56 subjects, with 30 of them in the control group and 26 in the test group. These subjects were exposed to three different tasks designed to induce varying stress levels: a relaxation mode (T1), a speech task (T2), and an arithmetic task (T3), each lasting for 3 min. The subjects were systematically divided into two groups: the control group and the test group. The control group was given less challenging tasks, whereas the test group faced more demanding tasks, serving to differentiate the stress levels induced in each group.

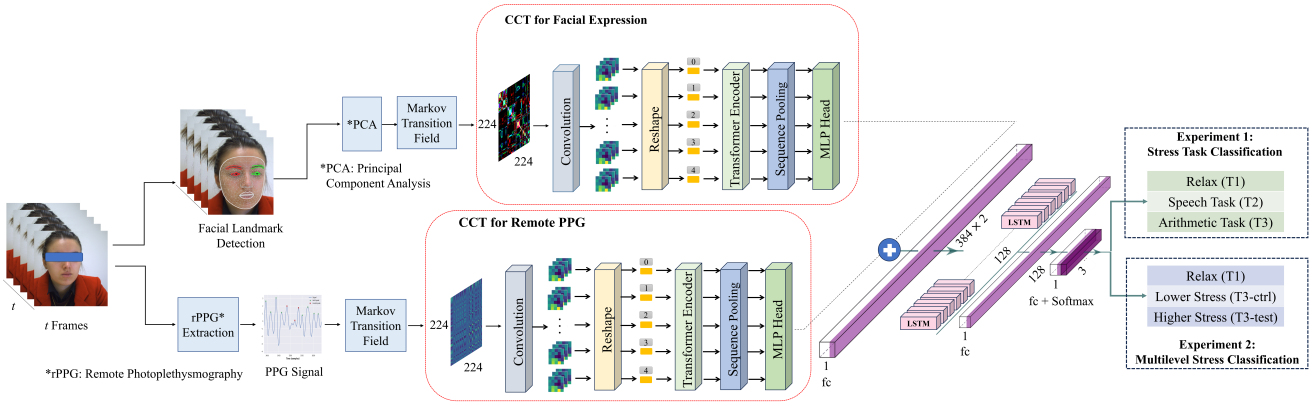

Figure 2. Schematic diagram of the proposed CCT-LSTM multimodal deep learning

4.2. Data Preprocessing for Facial Expression

To extract facial landmarks, we used the MediaPipe Face Landmarker [12], which is capable of identifying and distinguishing 478 landmarks. This method generates three-dimensional, normalized facial landmarks in the form of x, y, and z coordinates.

A window size of 60 s is applied with an overlapping moving interval of 5 s, then 478 landmarks are obtained from each frame in the window block. Following the landmark detection in the window block, a three-dimensional array is procured for each discrete window block. As presented in Figure 3, the three dimensions of this array correspond to the total number of frames (60 s $\times$ 35 fps), number of landmarks (478), and coordinates (3).

In the next step, the obtained array is decomposed to each coordinate to form three arrays with the shape of (total number of frames, number of landmarks). Next, to effectively compress the first dimension (total number of frames), a principal component analysis is employed for each coordinate to form a singular component array with the form of (1, number of landmarks). Following this step, we constructed an image representation using the MTF for each dimension.

As there are three dimensions, these are respectively assigned to the three color channels of an RGB image representation. This method thus permits the generation of a comprehensive and detailed image representation of facial landmarks while preserving the key characteristics and features of the original data.

4.3. Data Preprocessing for rPPG

The rPPG extraction method in our work is based on Face2PPG [4], which has three extraction processes: Rigid Mesh Normalization, Dynamic Selection of Facial Regions, and RGB to rPPG Conversion using Orthogonal Matrix Image Transformation (OMIT). Rigid Mesh Normalization

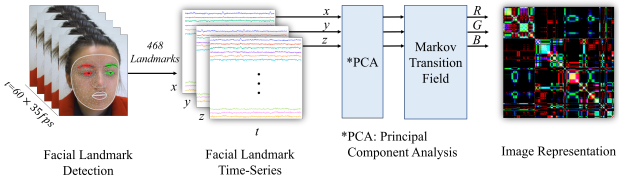

Figure 3. Data preprocessing for facial expression

stabilizes the detected face by normalizing the facial mesh, ensuring consistent signal extraction from the same facial location regardless of the pose or movement. Dynamic Selection of Facial Regions uses statistical and fractal analyses to dynamically select the facial regions that provide the best raw signal, discarding noisy or artifact-prone regions. RGB to rPPG Conversion using OMIT is a novel conversion method based on QR decomposition, which increases the robustness of signal extraction against compression artifacts.

In parallel to the data preprocessing applied for facial expression analysis, a temporal window size of 60 s, with an overlapping interval of 5 s, was used for the extraction of the rPPG signal. Following this extraction process, an image representation was created for each window block using MTF.

4.4. Model Validation

For the purposes of our experiments, we adopted two approaches:

1. **Stress task classification (T1, T2, and T3)**

2. **Multilevel stress classification (T1, T3-test, and T3-ctrl)**

In the first experiment, we employed sevenfold cross-validation. Then, 56 subjects were randomly divided into

7 subsets: 6 subsets were used for training, and the remaining 1 subset was used for testing the performance of the model. For the second experiment, we chose to focus solely on the arithmetic task (T3) as it was considered as the most challenging task among the three, thus likely to induce the highest level of stress. Our experiment aimed to classify stress into three levels: stress experienced in T3-ctrl group (lower stress), T3-test group (high stress), and relaxation task (T1). Due to the imbalanced number of subjects in the control and test groups, we used a stratified fivefold cross-validation approach to maintain the ratio while splitting the test and control groups into five separate subsets. In this case, four subsets were used as the training set to train the model, and the remaining one subset was used as the testing set to evaluate the performance of the model.

For both experiments, we trained the CCT individually for each modality (CCT-rPPG and CCT-Facial Landmark) without LSTM for 100 epochs to demonstrate the maximum precision achievable using CCT for each individual modality. Next, to investigate the performance of integrating the two modalities to form a CCT-LSTM multimodal deep learning architecture, we used the best epoch weights from each modality to initialize our multimodal framework and train for 100 epochs. All these steps were performed for every individual fold to ensure that the test data remained unseen by the models.

5. Results and Discussion

We presented a novel, multimodal deep learning method for remote stress estimation, which combines two modalities (1) facial expression based on 478 landmarks and (2) rPPG signals derived from video data.

In our first experiment, we focused on the classification of stress tasks T1, T2, and T3. For this purpose, we randomly divided 56 subjects into 7 subsets. Six of these subsets served as the training set, whereas the remaining subset was used for testing our model's performance. As discussed in the previous section first we trained a CCT for each individual modality, and then we trained our multimodal CCT-LSTM pipeline over 100 epochs for different task combinations: T1 vs. T2, T1 vs. T3, and T1 vs. T2 vs. T3. Table 1 reports our stress task classification results in comparison to the methodologies that Zhang *et al*. used for their comparative study. These methodologies include (1) MLP [8], which utilizes two fully connected layers to extract features from PPG and EDA data and then combines these features for stress classification using an additional four fully connected layers; (2) LIT [8], which uses two CNN layers for PPG feature encoding and two CNN plus two LSTM layers for EDA feature extraction, followed by classification through three CNN and four fully connected layers; (3) DFAF [13], which uses both inter and intramodal attention mechanisms to exchange and match information within and between modalities, subsequently averaging and fusing these joint representations; (4) CAM [21], which applies cross-modal attention to establish correlations between modalities and then fuses these for classification; and (5) MFN [28], which uses self-attention to encode unimodal features before combining them for stress detection. As presented in Table 1, the obtained results indicate that our multimodal CCT-LSTM outperforms all other state-of-the-art methods in binary and three-class task classification including the proposed methods by Sabour *et al*. and Zhang *et al*.

Our second experiment focused on multilevel stress classification. Due to the unequal number of participants in the control and test groups, we used a stratified fivefold cross-validation method to divide the test and control groups into five subsets. Four subsets were used for training, and the remaining one for testing. We trained the CCT model individually for each modality; facial expression and rPPG without using LSTM, for 100 epochs. The objective was to determine the highest precision achievable by CCT for each modality individually. As presented in Table 2, across the folds, the mean accuracy was 60.6% for rPPG and 62.0% for the facial expression CCT model. These outcomes indicated that the use of each modality in isolation is not a reliable indicator of stress levels. Contrarily, our multimodal CCT-LSTM pipeline showed a significant increase in performance, achieving an average accuracy of 80.5% and an average F1 score of 80.4%. The considerable improvement in these metrics underscores the strength of a multimodal approach and the combined application of CCT and LSTM for stress estimation. Our proposed framework demonstrated superior performance over all methods used in [22] for stress estimation, highlighting the potential advantages of using an integrated multimodal approach in this domain.

Subsequently, as detailed in Table 3, the mean values of precision, recall, and F1 score were determined by averaging across all five folds of the cross-validation process. This calculation was performed for each class, namely T1 (Rest), T3-ctrl (lower stress), and T3-test (higher stress). As presented in Table 3, among these classes, T3-ctrl outperformed the others by achieving the highest mean precision and F1 score, with values of 0.872 and 0.839 respectively. Alternatively, the T3-test class had the lowest mean scores for these metrics, registering a mean precision of 0.746 and an F1 score of 0.761.

In Table 4, the mean values of Precision, Recall, and F1 score across five different folds for each class are presented, and calculated using our multimodal CCT-LSTM for multilevel stress classification. A close examination of Table 4 shows that the highest F1 score is observed in the fourth fold, registering a remarkable 98.6%, whereas the fifth fold exhibits the lowest F1 score, at 72.0%. The exceptionally

Table 1. Experiment 1, stress task classification: Comparative experimental results between our method and other state-of-the-art methods on the UBFC-Phys dataset for stress task classification. The values in the table are the mean values (± standard deviations) of the sevenfold cross-validation, and the best results are in bold.

Methods	T1 vs. T2		T1 vs. T3		T1 vs. T2 vs. T3	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
MLP [8]	0.709(±0.061)	0.706(±0.064)	0.599(±0.040)	0.587(±0.037)	0.440(±0.028)	0.434(±0.031)
LIT [8]	0.701(±0.063)	0.699(±0.066)	0.625(±0.024)	0.622(±0.025)	0.447(±0.027)	0.443(±0.031)
DFAF [13]	0.758(±0.035)	0.756(±0.035)	0.689(±0.042)	0.686(±0.043)	0.478(±0.034)	0.477(±0.036)
CAM [21]	0.726(±0.060)	0.722(±0.063)	0.650(±0.052)	0.645(±0.054)	0.494(±0.028)	0.487(±0.033)
MFN [28]	0.769(±0.035)	0.768(±0.035)	0.666(±0.071)	0.664(±0.071)	0.501(±0.038)	0.490(±0.043)
BCSA [30]	0.818(±0.063)	0.817(±0.063)	0.723(±0.039)	0.722(±0.039)	0.558(±0.052)	0.560(±0.051)
CCT-rPPG (Ours)	0.804(±0.043)	0.803(±0.043)	0.766(±0.037)	0.765(±0.055)	0.592(±0.042)	0.568(±0.051)
CCT-Facial Landmark (Ours)	0.965(±0.016)	0.965(±0.016)	0.865(±0.041)	0.865(±0.042)	0.755(±0.048)	0.749(±0.048)
Multimodal CCT-LSTM (Ours)	**0.981(±0.016)**	**0.981(±0.016)**	**0.924(±0.037)**	**0.924(±0.037)**	**0.832(±0.058)**	**0.834(±0.056)**

Table 2. Experiment 2, multilevel stress classification: The experimental results of our proposed method on the UBFC-Phys dataset for multilevel stress classification. The values in the table are the mean values (± standard deviations) of the fivefold cross-validation, and the best results are in bold.

Methods	T1 vs. T3 ctrl vs. T3 test	
	Accuracy	F1 Score
CCT-rPPG	0.606(±0.041)	0.605(±0.042)
CCT-Facial Landmark	0.620(±0.100)	0.611(±0.103)
Multimodal CCT-LSTM	**0.805(±0.094)**	**0.803(±0.095)**

Table 3. Multimodal CCT-LSTM evaluation metrics for multilevel stress classification. The values in the table are the mean values. (± standard deviations) of the fivefold cross-validation.

Class	Precision	Recall	F1 score
T1	0.827(±0.131)	0.800(±0.116)	0.809(±0.109)
T3 ctrl	0.872(±0.989)	0.819(±0.127)	0.839(±0.094)
T3 test	0.746(±0.118)	0.807(±0.177)	0.761(±0.119)

high F1 score achieved in the fourth fold indicates of the robust capabilities of our multimodal CCT-LSTM model. However, the variation in F1 score across the different folds suggests that the dataset may not have a sufficient number of samples, which could be a potential limitation affecting the performance of the model.

6. Conclusion

Stress is a widespread concern in modern society, with the potential to cause serious long-term physical and mental health complications. The importance of stress detection extends beyond healthcare to include areas such as monitoring driver conditions and human–robot interaction. Despite its importance, a high-accuracy method for stress task and level classifications still remains a considerable challenge. To address this, we have introduced a multimodal deep learning approach for remote stress estimation, using the CCT-LSTM model.

Table 4. Multimodal CCT-LSTM k-fold cross-validation results for multilevel stress classification. The best fold results are in bold.

K-folds	Performance Metrics			
	Accuracy	Precision	Recall	F1 score
Fold-1	0.776	0.779	0.776	0.762
Fold-2	0.746	0.757	0.749	0.749
Fold-3	0.800	0.834	0.809	0.802
Fold-4	**0.986**	**0.985**	**0.987**	**0.985**
Fold-5	0.720	0.721	0.723	0.719
5-Fold Mean	0.805	0.815	0.809	0.803

This study highlights the potential of multimodal deep learning by presenting a novel CCT-LSTM model that integrates facial expressions and rPPG for stress level estimation, achieving state-of-the-art results in both stress task and level classifications. The first experiment set new benchmarks in binary and multiclass stress task classification, outperforming competing methods. The second experiment for multilevel stress classification showed significant improvement using the multimodal CCT-LSTM pipeline, achieving an average accuracy of 80.5% and an F1 score of 80.4% across fivefold cross-validation, compared with moderate accuracy using standalone modalities. However, the variation in F1 scores across different folds indicates the need for a more extensive dataset. The model outperformed prior attempts on the UBFC-Phys dataset, setting a new performance benchmark for remote stress task and level classifications. This highlights the potential of multimodal deep learning and the CCT-LSTM model for remote noninvasive stress detection, with significant implications for health

monitoring and human–robot interaction efficiency.

In the future, we plan to further different aspects of our research. This includes an in-depth analysis of specific hyperparameters such as window size and moving interval during the data preprocessing stage. We also aim to investigate alternative techniques for converting time-series data into images. These efforts will help us refine our methodologies and potentially enhance the accuracy of our stress detection model.

Acknowledgement

This study was partially supported by JSPS KAKENHI Grant Number 23K11152.

References

[1] Ane Alberdi, Asier Aztiria, and Adrian Basarab. Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review. *Journal of Biomedical Informatics*, 59:49–75, 2016. 1

[2] Shaibal Barua, Mobyen Ahmed, and Shahina Begum. Classifying drivers' cognitive load using eeg signals. *Studies in health technology and informatics*, 237:99–106, 6 2017. 1

[3] Frédéric Bousefsaf, Choubeila Maaoui, and Alain Pruski. Remote assessment of the heart rate variability to detect mental stress. pages 348–351, 2013. 2

[4] Constantino Álvarez Casado and Miguel Bordallo L'opez. Face2ppg: An unsupervised pipeline for blood volume pulse extraction from faces. *ArXiv*, abs/2202.04101, 2022. 5

[5] Weixuan Chen and Daniel McDuff. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pages 349–365, 2018. 1

[6] Nadezhda Chirkova and Sergey Troshin. Empirical study of transformers for source code. pages 703–715. Association for Computing Machinery, 2021. 3

[7] Gerard de Haan and Vincent Jeanne. Robust pulse rate from chrominance-based rppg. *IEEE Transactions on Biomedical Engineering*, 60:2878–2886, 2013. 1

[8] Tenzing C. Dolmans, Mannes Poel, Jan-Willem J.R. van 't Klooster, and Bernard P. Veldkamp. Perceived mental workload classification using intermediate fusion multimodal deep learning. *Frontiers in Human Neuroscience*, 14, jan 2021. 6, 7

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[10] Sami Elzeiny and Marwa Qaraqe. Blueprint to workplace stress detection approaches. pages 407–412, 2018. 1

[11] A. P. Association et al. Stress in america™ 2020: A national mental health crisis, 2020. 1

[12] Google for Developers. Mediapipe:face landmarker, 2022. https://developers.google.com/mediapipe/solutions/vision/face_landmarker. 5

[13] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. pages 6632–6641, 06 2019. 6, 7

[14] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, PP:1, 6 2019. 1

[15] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. *arXiv preprint arXiv:2104.05704*, 2021. 3, 4

[16] Lin He, Jiachen Ma, Sheikh Iqbal Ahamed, and Piyush Saxena. Quantitative multidimensional stress assessment from facial videos using deep learning. pages 710–715, 2022. 1

[17] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. Cstress: Towards a gold standard for continuous stress assessment in the mobile environment. pages 493–504. Association for Computing Machinery, 2015. 1

[18] Michele M Larzelere and Glenn N Jones. Stress and health. *Prim Care*, 35:839–856, 12 2008. 1

[19] Xin Liu, Josh Fromm, Shwetak Patel, and Daniel McDuff. Multi-task temporal shift attention networks for on-device contactless vitals measurement, 6 2020. 1

[20] Ryota Mitsuhashi, Kaito Iuchi, Takashi Goto, Akira Matsubara, Takahiro Hirayama, Hideki Hashizume, and Norimichi Tsumura. Video-based stress level measurement using imaging photoplethysmography. pages 90–95, 2019. 2

[21] Gnana Praveen R, Eric Granger, and Patrick Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition, 2021. 6, 7

[22] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. Ubfc-phys: A multimodal database for psychophysiological studies of social stress. *IEEE Transactions on Affective Computing*, 14:622–636, 2023. 1, 2, 4, 6

[23] Li Shen and Yangzhu Wang. Tcct: Tightly-coupled convolutional transformer on time series forecasting. *Neurocomputing*, 480:131–145, 2022. 3

[24] Zhaodong Sun, Alexander Vedernikov, Virpi-Liisa Kykyri, Mikko Pohjola, Miriam Nokia, and Xiaobai Li. Estimating stress in online meetings by remote physiological signal and behavioral features. pages 216–220. Association for Computing Machinery, 2023. 1

[25] Valeria Villani, Massimiliano Righi, Lorenzo Sabattini, and Cristian Secchi. Wearable devices for the assessment of cognitive effort for human–robot interaction. *IEEE Sensors Journal*, 20:13047–13056, 2020. 1

[26] Wenjin Wang, Albertus C den Brinker, Sander Stuijk, and Gerard de Haan. Algorithmic principles of remote ppg. *IEEE Transactions on Biomedical Engineering*, 64:1479–1491, 2017. 1

[27] Zhiguang Wang and Tim Oates. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. 6 2015. 4

[28] Han Yu, Thomas Vaessen, Inez Myin-Germeys, and Akane Sano. Modality fusion network and personalized attention in momentary stress detection in the wild. In *2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021*, 2021 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021, United States, 2021. Institute of Electrical and Electronics Engineers Inc. Funding Information: This work is supported by NSF 2047296 and 1840167. Publisher Copyright: © 2021 IEEE.; 9th International Conference on Affective Computing and Intelligent Interaction, ACII 2021 ; Conference date: 28-09-2021 Through 01-10-2021. 6, 7

[29] Zitong Yu, Xiaobai Li, and Guoying Zhao. Facial-video-based physiological signal measurement: Recent advances and affective applications. *IEEE Signal Processing Magazine*, 38:50–58, 2021. 1

[30] Xiaowei Zhang, Xiangyu Wei, Zhongyi Zhou, Qiqi Zhao, Sipo Zhang, Yikun Yang, Rui Li, and Bin Hu. Dynamic alignment and fusion of multimodal physiological patterns for stress recognition. *IEEE Transactions on Affective Computing*, pages 1–12, 2023. 2, 3, 7