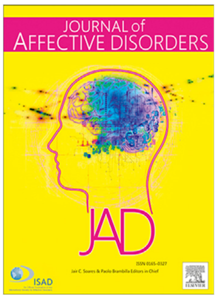
Journal Pre-proof



Task-aware multiple instance learning for stress detection from facial video data

Nele Sophie Brügge, Alexandra Korda, Heinz Handels,
Giorgos Giannakakis

PII: S0165-0327(26)00323-X
DOI: https://doi.org/10.1016/j.jad.2026.121472
Reference: JAD 121472

To appear in: *Journal of Affective Disorders*

Received date : 8 December 2025
Revised date : 13 February 2026
Accepted date : 19 February 2026

Please cite this article as: N.S. Brügge, A. Korda, H. Handels et al., Task-aware multiple instance learning for stress detection from facial video data. *Journal of Affective Disorders* (2026), doi: https://doi.org/10.1016/j.jad.2026.121472.

Title Page (with Author Details)

Task-Aware Multiple Instance Learning for Stress Detection from Facial Video Data

Nele Sophie Brügge[a,], Alexandra Korda[b], Heinz Handels[a,c,], Giorgos Giannakakis[d,e,f,]

[a] *AI in Medical Image and Signal Processing, German Research Center for Artificial Intelligence, Ratzeburger Allee 160, Lübeck, 23562, Germany*

[b] *Translational Psychiatry, Department of Psychiatry and Psychotherapy, University of Luebeck, Ratzeburger Allee 160, Lübeck, 23562, Germany*

[c] *Institute of Medical Informatics, University of Luebeck, Ratzeburger Allee 160, Lübeck, 23562, Germany*

[d] *Department of Electronic Engineering, Hellenic Mediterranean University, Romanou 3, Chania, 73133, Greece*

[e] *Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), N. Plastira 100, Heraklion, 70013, Greece*

[f] *Institute of Agri-food and Life Sciences, University Research and Innovation Center, Hellenic Mediterranean University, Heraklion, 71003, Greece*

Email addresses: `nele.bruegge@dfki.de` (Nele Sophie Brügge), `alexandra.korda@uni-luebeck.de` (Alexandra Korda), `heinz.handels@dfki.de` (Heinz Handels), `ggian@ics.forth.gr` (Giorgos Giannakakis)

Task-Aware Multiple Instance Learning for Stress Detection from Facial Video Data

Nele Sophie Brügge[a,*], Alexandra Korda[b], Heinz Handels[a,c,**], Giorgos Giannakakis[d,e,f,**]

[a] AI in Medical Image and Signal Processing, German Research Center for Artificial Intelligence, Ratzeburger Allee 160, Lübeck, 23562, Germany
[b] Translational Psychiatry, Department of Psychiatry and Psychotherapy, University of Luebeck, Ratzeburger Allee 160, Lübeck, 23562, Germany
[c] Institute of Medical Informatics, University of Luebeck, Ratzeburger Allee 160, Lübeck, 23562, Germany
[d] Department of Electronic Engineering, Hellenic Mediterranean University, Romanou 3, Chania, 73133, Greece
[e] Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), N. Plastira 100, Heraklion, 70013, Greece
[f] Institute of Agri-food and Life Sciences, University Research and Innovation Center, Hellenic Mediterranean University, Heraklion, 71003, Greece

Abstract

Stress is a prevalent condition linked to a wide range of mental and physical health disorders. Timely detection of stress is critical for enabling early interventions and promoting long-term well-being. Traditional detection methods based on physiological signals, can be effective, but intrusive or unsuitable for continuous or large-scale deployment. In this study, we propose a video-based stress detection method using top-k Multiple Instance Learning. Our approach is based on the assumption that subjects exhibit a mix of high-intensity and less stress-indicative behaviour during stressful tasks. We employ a temporal feature network with multi-head attention and introduce a conditioning mechanism to account for differences between active (speaking) and passive (non-speaking) tasks. To make better use of limited

[*] Corresponding author
[**] Shared last authorship

Email addresses: nele.bruegge@dfki.de (Nele Sophie Brügge), alexandra.korda@uni-luebeck.de (Alexandra Korda), heinz.handels@dfki.de (Heinz Handels), ggian@ics.forth.gr (Giorgos Giannakakis)

datasets and weak labels, we incorporate both top-k and bottom-k instances, assuming that bottom-k snippets reflect neutral or less stress-indicative behaviour even in stress-related videos. We validate our approach on our own stress dataset and the publicly available StressID dataset. In a leave-five-subjects-out evaluation, our method achieves high accuracy and F1 scores, outperforming baseline methods while providing interpretable temporal localisation of stress-related behaviour.

Keywords: Stress Detection, Multiple Instance Learning, Facial Expression Analysis, Weak Supervision, Affective Computing, Computer Vision

1. INTRODUCTION

Stress is a psychological response to situations perceived as threatening or challenging. While moderate stress can be beneficial, excessive or chronic stress can lead to negative impacts on physical and mental health [1]. Recognising stress early is especially critical for professionals in high-stakes or safety-critical jobs such as surgeons, pilots, long-distance drivers, or military personnel, and in individuals with existing mental health conditions, such as anxiety disorders. It would enable interventions that protect not only immediate functioning but also long-term health – reducing the risk of burnout, cognitive decline, and stress-related illnesses. Yet, automated stress detection is challenging and typically relies on biomarkers (such as cortisol, CRF, ACTH) or biosignals (such as ECG, EDA, and respiration) [2]. Collecting these signals often requires sensors that may be invasive or uncomfortable, potentially influencing the subject's stress response and complicating real-world monitoring.

In recent years, stress detection using facial features has gained growing attention as an alternative to biosignal-based methods. While it offers a convenient and non-invasive alternative, its performance often remains below that of biosignal-based approaches. For a more objective facial stress recognition, there has been an effort for identifying involuntary or semi-voluntary facial parameters [3, 4, 5, 6, 7] such as blinks, mouth micro-activity, or micro-expressions.

Still, the manifestation of stress in facial expressions is not yet fully understood and can vary significantly between individuals in both intensity and type. As a result, stress detection remains a challenging medical task for which creating fine-grained labelled datasets is difficult. Furthermore, annotating such data requires expert knowledge and is both time-consuming and labour-intensive, especially for videos. With weakly or noisily labelled data, supervised training of accurate classifiers becomes even more challenging, especially when only a few subtle signs of stress are present in a recording.

To address the challenges of limited annotation and subtle visual signals in facial video data, we propose using Multiple Instance Learning (MIL) for stress detection. MIL is well-suited for settings with weak, video-level labels, where fine-grained annotation is impractical. It enables models to learn from sets of instances (snippets) while identifying short, subtle segments of interest within longer recordings. In the MIL framework, videos that contain at least one target segment of interest are labelled as positive, while other videos are labelled as negative. A key assumption is that positive videos contain both positive and negative instances, while negative videos contain only negative instances. MIL has been successfully applied in anomaly detection from surveillance videos [8, 9, 10, 11, 12, 13] on datasets like (ShanghaiTech [14, 15], UCF-Crime [8], XD-Violence [16] and UCSD-Peds [17]).

For stress detection, we consider MIL appropriate due to the temporal sparsity of visual stress cues: even during stressful tasks, subjects often maintain a neutral expression for many frames and show a stress-related expression only briefly. Further, it does not only allow robust video-level classification but also enables snippet-level predictions, offering insight into when stress expressions occur. Our approach builds on top-k MIL [19, 11], which assumes that the k most confident snippets in a positive video are representative of the target class. Using these top-k snippet features, a classifier is trained to recognise stress-relevant patterns, which can be used to classify new snippets, and not only whole videos.

In this work, we extend our previously proposed MIL framework for stress detection [20], which incorporates an appropriate facial feature extraction method, the combined use of bottom-k and top-k snippets during training, and a tailored temporal attention mechanism combined with a bag-level classification network for binary classification. In [20], we introduced the use of bottom-k (lowest model confidence) snippets based on the assumption that they likely represent neutral or less stress-indicative facial expressions, even in videos globally labelled as containing stress, e.g., due to emotional inhibition [18]. While top-k snippets help the model focus on the most salient stress indicators, bottom-k snippets, presumed to reflect neutral behaviour, are explicitly assigned the neutral label to better utilise the available weak labels. Facial features are extracted using a contrastive learning network pretrained on temporal sequences of facial landmarks. For temporal modelling, we use a multi-head attention network designed to capture both short- and long-term expression patterns. In this work, we enhance the architecture with task-aware conditioning that differentiates between active (speaking) and passive (non-speaking) tasks, and evaluate multiple conditioning strategies to account for speech-related facial motion that can obscure stress- or emotion-related facial cues. In summary, our modifications to top-k MIL for stress detection consist of

- A contrastive learning-based facial feature extractor trained on temporal landmark sequences.
- A training scheme that leverages both top-k and bottom-k video snippets.
- A temporal feature extraction module based on multi-head attention.
- Speech-aware conditioning to distinguish between speaking and non-speaking conditions.

To validate our approach, we evaluate it on two datasets: our stress dataset, which comprises a set of controlled stress-inducing scenarios, and the publicly available StressID dataset [21].

2. Related Work

2.1. Stress Detection using Machine Learning

Recent research in stress detection using machine learning spans a wide range of methods. Traditional approaches, such as Random Forests and Support Vector Machines, have been applied successfully across various datasets and modalities [22, 23, 24, 25, 26, 27]. These studies frequently rely on physiological data collected from wearables, including electrodermal activity, electrocardiography, electroencephalography, and temperature sensors [22, 23, 24, 26, 25, 28, 29].

In parallel, video-based stress detection has gained attention as a non-invasive alternative [27, 29, 30, 31, 32], which greatly benefited from advancements in complex neural network architectures, achieving high accuracy in facial stress recognition [28, 29, 30, 31, 32, 33]. For example, Hasani and Mahoor [33] achieved up to 90 % accuracy using 3D convolutional networks for facial stress recognition. Other approaches integrate temporal information through temporal attention mechanisms [32] or LSTM layers [30, 31], often building on ResNet or I3D-based architectures [29]. Instead of using raw video input, some studies extract facial action units as input for classification models [7, 27, 34, 35]. Multimodal approaches that incorporate audio or ECG data alongside video have also been proposed, for example in [29], which yielded an accuracy of 85.1 % on data from 20 participants.

Stress detection has also been explored across diverse scenarios, including office environments [22], hospitals [26], driving [24], and social media [36]. Commonly used stress-inducing protocols involve cognitive tasks (e.g., mental arithmetic, memory recall), sensory stimuli (e.g., loud noises, emotionally arousing images), or social stressors. Despite extensive work in this field, MIL remains underexplored in stress detection. Building on prior work [20], we further investigate the effectiveness of MIL for video-based stress detection across different experimental tasks and stimuli from two video stress datasets.

2.2. Multiple Instance Learning in Medical Image and Video Analysis

MIL has shown strong potential across a range of medical image and video analysis tasks. In medical imaging, MIL has been used for dementia classification in brain MRI [37], diabetic retinopathy detection in colour fundus images [38] and hotspot detection in bone scintigraphy images [39]. It has also been widely applied to histopathology, particularly for cancer detection. Examples include identifying lymph node metastases in breast cancer [40, 41, 42], as well as classifying esophageal [38, 41] and colon cancers [43, 44, 45].

There is also work on medical video analysis, although MIL is more commonly applied to anomaly detection in surveillance camera videos [8, 9, 10, 11, 12, 13]. Sikka et al. [46, 47] employed MIL for pain localisation in medical videos. MIL has also been used for depression detection from facial landmarks [48], and for polyp frame detection in colonoscopy videos using a contrastive transformer-based approach [49].

3. Datasets

3.1. Our Stress Dataset

As a first dataset, we use our video stress dataset, comprising recordings of participants performing a series of stress-inducing and neutral tasks (dataset is described in [50]). The experimental protocol was designed to capture and investigate facial and physiological responses under stress conditions. The dataset includes videos of 58 participants (24 men, 34 women) with an average age of 26.9 ± 4.8 years.

Video acquisition protocol Participants were seated in front of a monitor and a camera. The camera's field of view covered the participant's face, with possible movements taken into account. It was mounted on a tripod and positioned at a distance of approximately 90 cm from the face. Ambient lighting conditions reduced the effects of specular lighting. Videos were recorded at a sampling rate of 60 frames per second and a resolution of 1216×1600 pixels, and later downsampled to 608×800 pixels at 30 frames per second.

Experimental tasks The experiment included eleven tasks: four neutral tasks, one relaxing task, and six stressful tasks in which stress conditions were simulated and induced using different types of stressors. Stress was induced across four distinct phases: *social exposure*, *emotional recall*, *mental workload tasks*, *stressful videos presentation*. The experimental tasks and their corresponding induced affective states are presented in Table 1. Every experiment began with a neutral or relaxing phase at each stage as a baseline, and each recording had a duration of 2 min.

The *social exposure* phase involved a brief interview in which participants were asked to describe themselves, a situation intended to simulate the stress of public exposure, similar to what an actor might face on stage. As a reference task, participants recited conventional sequences, such as counting from one to ten or listing the months of the year. In the *emotional recall* phase, stress was elicited by asking participants to recall and mentally relive a personally stressful past event, as if it were happening in the present. The *mental tasks* phase assessed cognitive load using two tasks: the modified Stroop Colour-Word Task (SCWT) [51], in which participants are asked to read colour names printed in incongruent ink (e.g., the word "RED" printed in blue), with increasing difficulty by alternating between reading and naming the ink colour; and the Paced Auditory Serial Addition Test (PASAT) [52], a neuropsychological task involving continuous mental arithmetic to evaluate attentional capacity under time pressure. Finally, the *stressful video* phase presented participants with 2-minute video clips designed to elicit emotional responses. These included both calming scenes and stress-inducing content, such as action sequences, scenarios involving heights (for participants with mild acrophobia), home invasions, and car accidents.

Table 1: Experimental tasks employed in our and the StressID dataset. The intended affective states of the experimental tasks are neutral (N), stress (S), and relaxed (R).

#	ID	Experimental Task	Affective State	Duration (min)
Our Dataset				
Social Exposure				
1	S1	Neutral (Reference)	N	2
2	S2	Baseline Description	N	2
3	S3	Interview	S	2
Emotional Recall				
4	E1	Neutral (Reference)	N	2
5	E2	Recall stressful event	S	2
Mental Workload				
6	M1	Reading words (Reference)	N	2
7	M2	Stroop Colour-Word Test	S	2
8	M3	PASAT task	S	2
Stressful Stimuli				
9	St1	Relaxing video	R	2
10	St2	Adventure video	S	2
11	St3	Psychological pressure video	S	2
StressID Dataset				
Baseline				
12	B1	Breathing: Guided breathing	R	3
13	B2	Baseline Speaking: Counting forward	N	1
14	B3	Relax: Calming video	R	5
Sensory Stimulation				
15	St1	Video 1: Comedy clip	R	1.5
16	St2	Video 2: Action scene	S	2
Mental Load				
17	M1	Counting 1: Backward -3 from 100	S	1
18	M2	Counting 2: Backward -7 from 1011	S	1
19	M3	Counting 3: Backward -3 from 1152 and finger tapping	S	1
20	M4	Stroop: Stroop Colour-Word Test	S	1
21	M5	Math: Solve 20 arithmetic problems	S	1
Social Exposure				
22	S1	Reading: Silent reading + summary	S	1
23	S2	Speaking: Describe strengths and weaknesses	S	1

3.2. StressID Dataset

As an additional data source, we used the publicly available StressID dataset [21]. The StressID dataset is a multimodal dataset for studying stress through synchronized physiological, audio, and video recordings. It includes data from 65 participants (47 men and 18 women) with an average age of 29 ± 7 years. Modalities include ECG, EDA, and respiration (all 65 subjects), facial video (54 subjects), and audio (56 subjects), totalling 39 hours of data.

Video acquisition protocol All participants were recorded using a Logitech QuickCam Pro 9000 RGB camera, positioned approximately 50 cm from the participant's face. The video data was captured at 720p resolution and 5 frames per second, while audio was recorded at 32 kHz with 16-bit resolution.

The StressID protocol comprises a structured series of tasks across three main categories: *sensory stimulation, mental workload*, and *psychosocial stress*. Each task is paired with self-reported labels for stress, relaxation, arousal, and valence. The tasks are detailed in Table 1. *Sensory stimulation* tasks, such as Video1 and Video2, involve passive viewing of emotionally charged video clips to activate the audiovisual cortex. Specifically, Video1 (from There's Something About Mary) targets low arousal and positive valence, while Video2 (from Indiana Jones and the Last Crusade) induces high arousal and negative valence. *Mental load* tasks include Counting1, Counting2, Math, Stroop, Reading, and Counting3, all of which demand mental effort through arithmetic, attention, comprehension, or multitasking, as Counting3 introduces dual-task interference by combining backward counting with coordinated hand movements. Psychosocial stress is elicited via the Speaking task, where participants discuss personal traits under evaluative pressure. The protocol begins with a relaxing breathing exercise to establish a physiological baseline, and concludes with a 5-minute calming video (Relax). Additionally, for the interactive stressors, a baseline was recorded, in which the participants were asked to count forwards.

4. Methods

4.1. Contrastive Learning Feature Extraction

Most MIL models rely on standard video feature extractors such as C3D [53] or I3D [54], which are typically pretrained on action recognition datasets like

Kinetics-400. However, these networks are not optimised for detecting subtle medical abnormalities in facial video data. We instead employ a contrastive learning network trained on landmarks extracted from facial video data. As demonstrated in [55], this network can extract discriminative features relevant to medical tasks. To preserve temporal coherence while promoting appearance invariance, the same landmark transformation is applied uniformly across all frames in each video snippet.

4.2. Multiple Instance Learning

4.2.1. Motivation

Multiple Instance Learning is a weakly supervised approach that uses video-level labels to infer snippet-level predictions. The video data is treated as bags of shorter video snippets: in our stress detection task, positive bags contain at least one snippet showing stress behaviour, and negative bags represent videos showing neutral or relaxed behaviour only. Top-k MIL [56], [57], [11] uses the top-k most likely positive instances within each bag to classify both on instance- and bag level. We evaluate the classification performance at the bag level due to the absence of second-wise labelled data. Simultaneously, MIL provides an interpretable temporal segmentation.

We assume that stress behaviour only appears in a fraction of snippets, while most appear neutral or at least less stress-indicative. To effectively use the available data, we consider not only top-k snippets but also bottom-k snippets within positive bags for training the network. Building on our assumption, we label these bottom-k snippets as neutral. Additionally, we introduce a task-aware conditioning network to help the network differentiate between passive and active tasks as speech-related facial motion could obscure emotion or stress-related cues.

4.2.2. Conditional Bottom-k Multiple Instance Learning

For stress detection, videos are segmented into shorter temporal snippets, which are grouped into fixed-size bags. For each snippet, features are extracted using the contrastive learning network described in subsection 4.1. The result is a set of video-level features $\mathbf{F}_i \in \mathbb{R}^{T \times D}$ and corresponding binary stress labels y_i, forming a weakly labelled dataset $\mathcal{D} = (\mathbf{F}_i, y_i)_{i=1}^{|\mathcal{D}|}$, where D is the feature dimensionality and T the number of snippets. Labels are assigned at the video level: $y_i = 0$ for neutral or relaxed states (N or R), and $y_i = 1$ for stress (S).

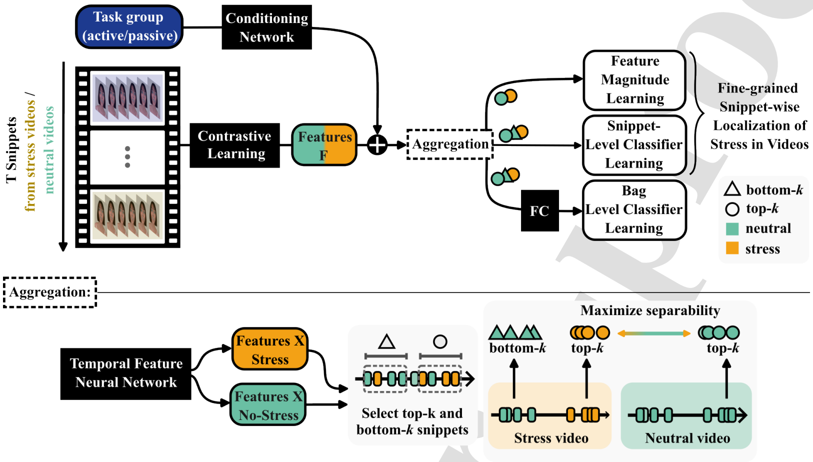


Figure 1: The proposed conditional multiple instance learning approach for stress detection. We divide each video into T snippets, from which we extract contrastive learning features. A conditioning network embeds the task group label (0 for passive tasks, 1 for active tasks). This embedding is concatenated with the contrastive features and passed to a temporal feature extraction network. The separability of neutral and stress snippets is maximised by feature magnitude learning. Additionally, top-k and bottom-k snippets of stress videos, along with the top-k snippets of neutral videos, are used as input to a bag-level classifier and a snippet-level. Bottom-k snippets are treated as neutral, even if they originate from a stress video.

To condition the network on the type of task performed, we introduce an additional speech label $z_i \in \{0, 1\}$, where $z_i = 0$ denotes *passive* tasks (no speaking), and $z_i = 1$ denotes *active* tasks (involving speaking). This task group label is passed through a small conditioning network g_ω, consisting of a fully-connected layer that maps the scalar task label to a D^{cond}-dimensional vector:

$$\mathbf{c} = g_\omega(z_i) \in \mathbb{R}^{D^{\text{cond}}}. \quad (1)$$

We investigate three conditioning mechanisms: a linear embedding network, Feature-wise Linear Modulation (FiLM) [58], and cross-attention and apply them either before or after the temporal feature network s_θ described in subsection 4.3. In the *linear* variant, the task embedding $\mathbf{c}$ is concatenated to each temporal feature vector. Let $\mathbf{F} \in \mathbb{R}^{T\times D}$ denote the sequence of temporal features. The task embedding is repeated along the temporal axis, yielding $\mathbf{C} \in \mathbb{R}^{T \times D^{\text{cond}}}$, and concatenated as

$$\mathbf{F}^{\text{cond, linear}} = \text{Concat}(\mathbf{F}, \mathbf{C}) \in \mathbb{R}^{T \times (D + D^{\text{cond}})}. \tag{2}$$

This conditioning can be applied either before the temporal network s_θ or to its output features.

For FiLM, the task embedding $\mathbf{c}$ is mapped to a pair of modulation parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ via a learnable projection $(\boldsymbol{\gamma}, \boldsymbol{\beta}) = h_\phi(\mathbf{c})$ with $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^D$. The conditioning is then applied by modulating the feature activations as

$$\mathbf{F}^{\text{cond, FiLM}} = \mathbf{F} \odot (1 + \boldsymbol{\gamma}) + \boldsymbol{\beta}, \tag{3}$$

where $\odot$ denotes element-wise multiplication. When applied before the temporal network, modulation operates on the input feature sequence; when applied after, it modulates the output of s_θ.

In the cross-attention conditioning, the projected task embedding $\mathbf{q} = W_q\mathbf{c} \in \mathbb{R}^D$ is used to query the temporal feature sequence. The temporal features are projected to keys $\mathbf{K} = W_k\mathbf{F}$ and values $\mathbf{V} = W_v\mathbf{F}$. Cross-attention is then computed as $\mathbf{A} = \text{MultiHeadAttn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) \in \mathbb{R}^{T \times D}$. The task-conditioned feature refinements are integrated via a gated residual connection with the sigmoid function $\sigma(\cdot)$:

$$\mathbf{F}^{\text{cond, Cross-Att}} = \mathbf{F} + \sigma(W_g\mathbf{c}) \cdot \mathbf{A}. \tag{4}$$

As the other mechanisms, cross-attention conditioning was applied either before or after the temporal network s_θ.

Figure 1 shows an overview of our bottom-k MIL approach. Temporal dependencies are modelled using a multi-head attention temporal feature network $s_\theta : \mathcal{F} \to \mathcal{X}$ (see subsection 4.3 and Fig. 1 in the Supplementary Material for details). This network transforms the conditioned input features $\mathbf{F}^{\text{cond}}$ into temporal features $\mathbf{X} = s_\theta(\mathbf{F}^{\text{cond}})$. The resulting features are passed to a snippet-level classifier $f_\phi : \mathcal{X} \to [0, 1]^T$, which predicts whether a video snippet contains stress-related behaviour. Following [11], we denote the snippet-level features corresponding to stress and non-stress behaviour as $\mathbf{x}^+ \sim P_x^+(\mathbf{x})$ and $\mathbf{x}^- \sim P_x^-(\mathbf{x})$. With $t = 1, ..., T$, a snippet feature $\mathbf{x}_t$ represents the t-th row in $\mathbf{X}$. Stress-positive videos $\mathbf{X}^+$ can contain snippets drawn from both $P_x^+(\mathbf{x})$ and $P_x^-(\mathbf{x})$, but stress-negative videos $\mathbf{X}^-$ can only contain snippets from $P_x^-(\mathbf{x})$. We make the assumption that stress snippets produce features with higher magnitude than non-stress snippets, i.e. $\mathbb{E}\left[\|\mathbf{x}^+\|_2\right] \geq \mathbb{E}\left[\|\mathbf{x}^-\|_2\right]$.

The snippet-level classifier f_ϕ, the temporal feature network s_θ, and the bag-level classifier c_ψ are trained jointly using the following loss function:

$$\begin{aligned}\ell_{\text{overall}} = \min_{\phi,\theta,\psi} \sum_{i,j=1}^{|\mathcal{D}|} \sum_{n=1}^{N} &\ell_s(s_\theta(\mathbf{F}_i^{(n),\text{cond}}), s_\theta(\mathbf{F}_j^{(n),\text{cond}}), y_i, y_j) \\ &+\ell_f(f_\phi(s_\theta(\mathbf{F}_i^{(n),\text{cond}})), y_i) \\ &+\ell_b(c_\psi(f_\phi(s_\theta(\mathbf{F}_i^{(n),\text{cond}})), y_i)).\end{aligned} \tag{5}$$

Here, N denotes the number of sub-segments (bags) of length T extracted from each recording. The overall loss combines three components: the feature separability loss ℓ_s, the snippet-level classification loss ℓ_f, and the bag-level classification loss ℓ_b. We describe each loss term in detail below.

We adopt the feature separability loss ℓ_s from [11] to encourage higher feature magnitudes for stress-related snippets. The mean feature norm of the top-k snippets is calculated by

$$g_{\theta,k}(\mathbf{X}) = \max_{\Omega_k(\mathbf{X}) \subseteq \{\mathbf{x}_t\}_{t=1}^{T}} \frac{1}{k} \sum_{\mathbf{x}_t \in \Omega_k(\mathbf{X})} \|\mathbf{x}_t\|_2 \tag{6}$$

where $\Omega_k(\mathbf{X})$ is a subset of k snippets in $\{\mathbf{x}_t\}_{t=1}^{T}$. With a pre-defined margin m, the separability loss ℓ_s compares the top-k magnitudes between positive ($\mathbf{X}^+$) and negative ($\mathbf{X}^-$) samples and is given by

$$\begin{aligned}&\ell_s(s_\theta(\mathbf{F}_i^{\text{cond}}), s_\theta(\mathbf{F}_j^{\text{cond}}), y_i, y_j)) = \\ &\begin{cases}(|m - g_{\theta,k}(\mathbf{X}^+)| + g_{\theta,k}(\mathbf{X}^-))^2 & \text{if } y_i = 1, y_j = 0, \\ 0 & \text{otherwise.}\end{cases}\end{aligned} \tag{7}$$

The snippet-level classification cross-entropy loss l_f uses binary cross-entropy to supervise predictions at the snippet level and is defined by

$$\begin{aligned}&l_f(f_\phi(s_\theta(\mathbf{F}^{\text{cond}})), y) = \\ &\sum_{\substack{\mathbf{x} \in \Omega_{k,\max}(\mathbf{X}) \\ \mathbf{x} \in \Omega_{k_b,\min}(\mathbf{X})}} - \left(y' \log(f_\phi(\mathbf{x})) + (1 - y') \log(1 - f_\phi(\mathbf{x}))\right),\end{aligned} \tag{8}$$

where $\Omega_{k,\max}(\mathbf{X})$ denotes the set of top-k snippets with the highest L2 norms, and $\Omega_{k_b,\min}(\mathbf{X})$ the set of bottom-k_b snippets with the smallest L2 norms. For top-k snippets, the target label is the original bag label $y' = y$. For bottom-k snippets, we assign $y = 0$, regardless of the original bag label.

The bag-level classification loss ℓ_b supervises predictions made at the aggregated bag level. To form the bag-level prediction, features from both top-k and bottom-k snippets are combined into a feature set $\tilde{\mathbf{X}} = \Omega_{k,\max}(\mathbf{X}) \cup \Omega_{k_b,\min}(\mathbf{X})$, which are then passed to a simple bag-classification head $c_\psi(\cdot)$ consisting of two fully-connected layers and ReLU activation. The loss is defined as

$$\begin{aligned} &\ell_b(c_\psi(f_\phi(s_\theta(\mathbf{F}^{\text{cond}}))), y) = \\ &-\Big(y \log\Big(c_\psi(f_\phi(\tilde{\mathbf{X}}))\Big) + (1-y)\log\Big(1 - c_\psi(f_\phi(\tilde{\mathbf{X}}))\Big)\Big). \end{aligned} \tag{9}$$

4.3. Extracting Temporal Features

For temporal feature extraction, we use a Multi-Scale Multi-Head Attention Network (MSMHN) introduced in [20], which employs multi-head attention and convolutions at different temporal scales. An overview of the architecture is provided in the Supplementary Material in Fig. 1. The input to the network is a sequence of snippet-level features $\mathbf{F} \in \mathbb{R}^{T\times D}$ obtained from the contrastive learning encoder, concatenated with a conditioning feature vector $\mathbf{C} \in \mathbb{R}^{T\times D^{\text{cond}}}$, yielding $\mathbf{F}^{\text{cond}} \in \mathbb{R}^{T\times(D+D^{\text{cond}})}$. These conditioned features $\mathbf{F}^{\text{cond}}$ are then passed to the MSMHN for temporal encoding. The temporal features are extracted at different temporal scales using parallel 1D dilated convolutional branches, following [11].

Each branch is followed by a multi-head self-attention module [59], allowing the network to focus on relevant parts of the input, as also explored in other attention-based approaches [60, 61, 62]. The outputs of all branches are concatenated and further processed through a convolutional layer for feature fusion. This resulting temporal feature representation $\mathbf{X} = s_\theta(\mathbf{F}^{\text{cond}})$ is finally added to the input features via a skip connection.

5. Experimental Details

Experiments We trained the self-supervised contrastive feature network according to [55] and applied it to both our dataset and the StressID dataset. For applying the MIL model, we separate the data into bags of $T = 30$ (our dataset) or $T = 20$ (StressID dataset) consecutive feature snippets of one subject performing a single task. Bags were chosen without overlap in our dataset and overlapped by 15 s in the StressID dataset to increase the dataset size.

We conduct three training setups for our dataset. In the first experiment, we train the MIL framework on data from one neutral and one stressful video from the same experimental task, yielding a balanced dataset. This process is repeated for all task combinations, resulting in one network trained for each task pair. In the second experiment, we combine tasks into two groups: either all active tasks (where subjects speak) or all passive tasks (where subjects do not speak), training one shared model per group. Finally, we train a single model on all tasks jointly.

Using the StressID dataset, we again train a model on task pairs in a first experiment. For this, we use the baseline task in which subjects were asked to count forward and a second, active task. Since the StressID dataset additionally contains self-assessment labels, we also train networks using these labels instead of task labels. Again, we also train a model on all combined tasks in a second experiment and investigate the influence of conditioning on the results.

Hyperparameters and training details We trained the proposed MIL framework using the Adam optimiser with an initial learning rate of 10^{-4} and a batch size of 32. The top-k parameter was set to $k = 10$ with $k_b = k$ for bottom-k snippets. We included an ablation study on the hyperparameter k in Figure 2 in the Supplementary Material. The bag-level classifier c_ψ is implemented as a two-layer fully-connected neural network with 512 nodes in the hidden layer and a ReLU activation function. We trained the model for 10 epochs on our stress dataset and 60 epochs on the StressID dataset. When training on all tasks from our stress dataset jointly, the model was trained for 100 epochs. For the conditioning vector, we choose a dimensionality of 32, while an ablation study shows that performance remains stable across a wide range of embedding sizes (Supplementary Material, Table 2).

To evaluate generalisation performance, we performed a 10-fold cross-validation, where each fold holds out data from 5 subjects for evaluation. This subject-wise partitioning ensures that the model is tested on individuals it has not seen during training, providing an estimate of generalisation to unseen subjects.

6. Results

This section presents our results across different networks and training schemes. Ablation results are listed in Table 2. Additionally, we report the outcomes of the three experimental training setups: task-specific training, training on grouped active or passive tasks, and joint training on all tasks (see Table 3 and Table 4). In the following, we outline the experiments using different network architectures and training schemes. We trained all networks on the same contrastive learning features to ensure comparability.

Table 2: Stress classification bag-level accuracy (ACC) and F1 Score (F1) for different network architectures and learning schemes for our and StressID datasets. For our dataset, results were averaged over all task combinations. The column "MIL" indicates whether the MIL training scheme was used. "Dense Label" indicates that the classification was performed on a second-wise snippet basis where all snippets were assigned the bag label. "Bottom-*k*" means that we used the proposed MIL approach from subsection 4.2.2.

Model	MIL	Our Dataset		StressID	
		ACC (%)	F1 (%)	ACC (%)	F1 (%)
ResNet-18 + Dense Label	✗	77.28 ± 16.92	76.34 ± 23.45	72.78 ± 3.25	72.65 ± 3.25
ResNet-18	✗	83.38 ± 12.88	81.32 ± 21.43	74.84 ± 5.11	74.71 ± 4.86
MTN [11]	✗	86.14 ± 12.89	82.68 ± 12.43	75.51 ± 6.54	74.17 ± 6.96
MTN [11]	✓	93.19 ± 5.21	93.57 ± 4.71	76.62 ± 3.83	76.07 ± 3.78
MTN [11] + Bottom-*k*	✓	94.17 ± 5.19	94.28 ± 5.13	76.24 ± 4.75	75.01 ± 5.49
MSMHN	✓	95.09 ± 4.77	95.22 ± 4.63	75.73 ± 8.58	75.31 ± 8.37
MSMHN + Bottom-*k*	✓	$\mathbf{95.46 \pm 4.37}$	$\mathbf{95.49 \pm 4.77}$	$\mathbf{78.16 \pm 4.35}$	$\mathbf{78.12 \pm 3.93}$

3D ResNet-18 trained with dense labels As a baseline model, we trained a 3D ResNet-18 [63] in a fully supervised way on snippet level. To get a densely-labelled dataset, we labelled each snippet with the bag-label and individually fed these snippets to the ResNet. This resulted in an accuracy of 77.28 % and an F1 score of 76.34 %. For StressID, this baseline yielded $72.78 \pm 3.25\%$ accuracy and $72.65 \pm 3.25\%$ F1 score.

3D ResNet-18 trained with bag-level labels In a second experiment, we trained the same 3D ResNet-18 architecture but assigned a single label to each 30 s segment. Unlike the first experiment, where single snippets were classified individually, all 30 snippet features within a segment were fed to the model simultaneously. This segment-wise approach improved performance, reaching $83.38 \pm 12.88\%$ accuracy and $81.32 \pm 21.43\%$ F1 on our stress dataset, and $74.84 \pm 5.11\%$ accuracy with $74.71 \pm 4.86\%$ F1 on StressID.

Table 3: Stress classification bag-level accuracy (ACC) and F1 score (F1) for all stress task combinations from our stress dataset. As network we used the MSMHN and trained it using the proposed bottom-score MIL approach. In addition to individual tasks, we report performance for grouped passive (S1, E1, St1 vs. E2, St2, St3) and active (S2, M1 vs. S3, M2, M3) task sets, as well as for training on all tasks jointly. In the joint training setting, we further evaluated different conditioning mechanisms, where the network receives passive/active task labels (speech conditioning) or task category information (task type conditioning) as input. The conditioning type refers to the mechanism used (linear projection, FiLM or cross-attention), which can be applied to the snippet features (pre) or the temporal features (post).

	Cond.			
Task	Type	Pos.	ACC (%)	F1 (%)
No Conditioning				
Social Exposure (S2 vs. S3)	–	–	97.78 ± 1.57	97.78 ± 1.55
Emotional Recall (E1 vs. E2)	–	–	96.77 ± 2.80	96.82 ± 2.63
Mental Workload (M1 vs. M2)	–	–	94.03 ± 3.46	94.34 ± 3.18
Mental Workload (M1 vs. M3)	–	–	95.19 ± 3.14	95.35 ± 3.00
Stressful Stimuli (St1 vs. St2)	–	–	97.35 ± 2.08	97.29 ± 2.09
Stressful Stimuli (St1 vs. St3)	–	–	91.68 ± 7.15	91.35 ± 7.64
Passive Tasks	–	–	93.54 ± 3.03	93.39 ± 3.04
Active Tasks	–	–	95.32 ± 2.27	95.03 ± 2.53
All Tasks	–	–	62.51 ± 2.27	63.09 ± 2.21
Speech Conditioning				
All Tasks	Linear	pre	96.54 ± 1.46	96.41 ± 1.49
All Tasks	FiLM	pre	96.88 ± 1.08	96.75 ± 1.11
All Tasks	Cross-Att.	pre	96.17 ± 1.81	95.99 ± 1.90
All Tasks	Linear	post	96.26 ± 1.65	96.12 ± 1.69
All Tasks	FiLM	post	96.61 ± 1.44	96.48 ± 1.47
All Tasks	Cross-Att.	post	96.27 ± 1.36	96.10 ± 1.40
Task Conditioning				
All Tasks	Linear	pre	96.69 ± 1.76	96.55 ± 1.82
All Tasks	FiLM	pre	96.95 ± 1.28	96.82 ± 1.32
All Tasks	Cross-Att.	pre	96.67 ± 1.81	96.51 ± 1.74
All Tasks	Linear	post	96.51 ± 1.58	96.38 ± 1.62
All Tasks	FiLM	post	97.28 ± 1.56	97.18 ± 1.61
All Tasks	Cross-Att.	post	96.38 ± 1.66	96.21 ± 1.72

Table 4: Stress classification task and self-assessment label bag-level accuracy (ACC) and F1 score for 8 task combinations from the StressID dataset. As network we used the MSMHN architecture trained using the proposed bottom-score MIL approach. Further, in separate experiments, we evaluated the results on the combined tasks. When using all tasks, we additionally use conditioning (Cond.), where we evaluated different conditioning mechanisms on passive/active task labels (speech conditioning). The conditioning type refers to the mechanism used (linear projection, FiLM or cross-attention), which can be applied to the snippet features (pre) or the temporal features (post).

	Cond.		Task labels		Self-assessment	
Task	Type	Pos.	ACC (%)	F1 (%)	ACC (%)	F1 (%)
No Conditioning						
Mental Load (B2 vs. M1)	–	–	69.70 ± 7.23	69.30 ± 7.32	64.26 ± 13.12	64.91 ± 11.40
Mental Load (B2 vs. M2)	–	–	75.20 ± 12.56	75.60 ± 12.29	74.12 ± 8.89	73.25 ± 7.83
Mental Load (B2 vs. M3)	–	–	62.85 ± 4.87	60.58 ± 5.71	58.78 ± 13.20	60.92 ± 11.84
Mental Load (B2 vs. M4)	–	–	69.44 ± 8.19	68.29 ± 8.37	63.71 ± 9.57	64.41 ± 9.94
Mental Load (B2 vs. M5)	–	–	78.41 ± 9.62	77.92 ± 10.01	72.78 ± 9.58	70.16 ± 9.51
Sens. Stim. (St1 vs. St2)	–	–	51.60 ± 11.21	50.43 ± 10.71	57.71 ± 20.67	53.25 ± 12.73
Social Exp. (B2 vs. S1)	–	–	70.09 ± 7.31	68.58 ± 8.10	57.92 ± 10.39	61.73 ± 11.51
Social Exp. (B2 vs. S2)	–	–	71.39 ± 13.93	69.42 ± 12.43	60.60 ± 8.06	64.06 ± 8.53
All Tasks	–	–	78.16 ± 4.35	78.12 ± 3.93	67.92 ± 10.33	68.50 ± 10.14
Speech Conditioning						
All Tasks	Linear	pre	81.39 ± 0.87	83.72 ± 0.82	71.29 ± 10.20	71.95 ± 10.41
All Tasks	FiLM	pre	80.17 ± 2.94	80.17 ± 4.42	69.41 ± 9.21	69.41 ± 10.18
All Tasks	Cross-Att.	pre	78.32 ± 3.36	78.32 ± 3.76	61.72 ± 5.78	61.72 ± 10.52
All Tasks	Linear	post	81.38 ± 0.80	81.38 ± 0.70	68.86 ± 10.64	68.86 ± 11.61
All Tasks	FiLM	post	81.55 ± 0.87	81.55 ± 0.84	66.52 ± 10.13	66.52 ± 10.94
All Tasks	Cross-Att.	post	80.34 ± 1.73	80.34 ± 2.03	68.98 ± 9.80	68.98 ± 10.97

MTN We implemented the MTN architecture from [11], combining dilated convolutions, temporal attention, and a residual connection. Again, all extracted features were passed to a bag classifier. On our stress dataset, the MTN achieved $86.14 \pm 12.89\%$ accuracy and $82.68 \pm 12.43\%$ F1. On StressID, it resulted in $75.51 \pm 6.54\%$ accuracy and $74.17 \pm 6.96\%$ F1.

MTN trained using MIL The MTN was then trained using a top-k MIL approach without bottom-k features. This significantly improved the results on our stress dataset to $93.19 \pm 5.21\%$ accuracy and $93.57 \pm 4.71\%$ F1. On StressID, a modest gain was observed with $76.62 \pm 3.83\%$ accuracy and $76.07 \pm 3.78\%$ F1.

MTN trained using MIL with bottom scores Next, we included bottom-k features into the MIL training for both snippet- and bag-level classification losses ℓ_f and ℓ_b. On our stress dataset, performance improved slightly to $94.17{\pm}5.19\%$ accuracy and $94.28{\pm}5.13\%$ F1. However, for StressID, results slightly dropped to $76.24{\pm}4.75\%$ accuracy and $75.01{\pm}5.49\%$ F1.

MSMHN trained using MIL We trained the proposed MSMHN, from Section 4.3, using top-k MIL only. This model surpassed all previous results on our stress dataset, reaching $95.09 \pm 4.77\%$ accuracy and $95.22 \pm 4.63\%$ F1. On StressID, it achieved $75.73 \pm 8.58\%$ accuracy and $75.31 \pm 8.37\%$ F1.

MSMHN trained using MIL with bottom scores Finally, integrating bottom-k features in the MIL training of MSMHN yielded the best results overall with an accuracy of 95.46 % and an F1 score of 95.49 %, outperforming all previously considered methods. We conducted statistical analyses on the collected results. Normality of the data was verified using the Shapiro–Wilk test. A paired t-test revealed statistically significant differences in the performance measures compared to MSMHN without bottom scores. Specifically, the F1 scores yielded $t(59) = 2.324, P = 0.023$, significant at $P < 0.05$, and the accuracy yielded $t(59) = 3.751, P = 0.0004$, significant at $P < 0.001$. On StressID, it reached the highest performance with $78.16 \pm 4.35\%$ accuracy and $78.12 \pm 3.93\%$ F1 score.

Evaluation on single tasks and task combinations The performance of the best-performing model (MSMHN trained using MIL with bottom scores) across all task combinations in our stress dataset is summarised in Table 3. The table shows that the classification yielded the best results for the subjects performing a baseline description vs. being involved in an interview (S2 vs. S3) and watching a relaxing video vs. an adventure video (St1 vs. St2). In these tasks, the classification accuracies are 97.78 % and 97.35 % and the

F1 scores are 97.78 % and 97.29 %, respectively. In contrast, the most challenging setting corresponds to distinguishing between relaxing and watching psychological pressure videos (St1 vs. St3), for which the model achieves an accuracy of 91.68 % and an F1 score of 91.35 %.

Further, we group tasks into passive and active categories and train the model on all grouped tasks jointly. The classification performance remains high. Passive tasks (e.g., video watching and emotional recall) yield slightly lower performance than active tasks (e.g., social exposure and mental workload), with average accuracies of 93.54 % and 95.32 % and F1 scores of 93.39 % and 95.03 %, respectively.

For each task pair in the StressID dataset, we present the results in Table 4. Across tasks, performance varied depending on the type and intensity of the stressor. Using the *task labels* as in our stress dataset, the task combination B2 vs. M5 (Baseline vs. Math) yielded the best performance with an accuracy of 78.41 % and an F1 score of 77.92 %. In contrast, with an accuracy of 51.60 % and an F1 score of 50.43 %, the lowest performance was observed in the Sensory Stimulation setting, where the participants watched videos extracted from two different movies.

When evaluating performance based on *self-assessment labels*, a similar trend was observed, though the absolute values were slightly lower overall. The best performance again occurred in the B2 vs. M2 task (accuracy: 74.12 %, F1: 73.25 %), while the sensory stimulation task (St1 vs. St2) showed the lowest classification results (accuracy: 57.71 %, F1: 53.25 %).

Modelling tasks jointly using conditioning In another experiment, we jointly modelled both speech and non-speech tasks and analysed the effect of conditioning, with quantitative results reported in Tables 3 and 4. We evaluate three conditioning mechanisms: linear projection, FiLM [58], and cross-attention, applied either before (pre-temporal) or after (post-temporal) the temporal network described in subsection 4.3. Across all configurations, performance differences between conditioning mechanisms and insertion points are relatively small, indicating that the choice of conditioning architecture only has a limited impact on overall performance. We compare two types of conditioning signals: task type (differentiating between Social Exposure, Emotional Recall, Mental Workload and Stressful Stimuli) and speech activity. Experiments are conducted on both datasets; however, for StressID, only speech-based conditioning is considered, as task identity would introduce a strong prior due to the tight coupling between task type and stress labels. An overview of all conditioning variants is shown in the Supplementary Material in Table 1.

On our stress dataset, training jointly across all tasks without conditioning leads to a substantial performance drop, with an accuracy of 62.51 % and an F1 score of 63.09 %. Introducing conditioning markedly improves performance across all variants, where task- and speech-based conditioning achieve comparable performance. FiLM-based conditioning yields the best results, with post-temporal task type conditioning achieving the highest F1 score (97.18 %), closely followed by pre-temporal task type conditioning (96.82 %).

On StressID with self-assessment labels, jointly training on all tasks without conditioning yields an accuracy of 67.92 % and an F1 score of 68.50 %. Incorporating speech-based conditioning results in consistent improvements across most configurations, with the best performance obtained using linear conditioning before the temporal network, achieving 71.29 % accuracy and 71.95 % F1. For the task-label setting on StressID, joint training across tasks without conditioning achieves an accuracy of 78.16 % and an F1 score of 78.12 %. Adding speech-based conditioning further improves performance, with linear pre-temporal conditioning achieving the strongest results at 81.39 % accuracy and 83.72 % F1.

Generalisation to unseen tasks To assess robustness, we evaluate the model under task hold-out settings, where entire tasks are excluded during training. Performance remains largely stable, with only marginal drops compared to training on all tasks, with F1 scores ranging from 82.24 % (Emotional Recall) to 98.85 % (Social Exposure). Even when multiple tasks are jointly held out (e.g., S2, S3, E1, E2), performance decreases only slightly to 91.20 %. Full results for the task exclusion setting are provided in the Supplementary Material in Table 3.

Explainability Through Snippet-Level Predictions We present exemplary sequences of facial feature magnitudes and corresponding snippet-level classifier scores for five subjects during relaxed (task 4.1) and stressed (task 4.3) tasks in Figure 2. The plots show the feature magnitudes and snippet classifier scores over time, highlighting the time steps the network focuses on, enhancing the interpretability of our model and offering insights into the temporal dynamics of stress expressed in facial behaviour. The results indicate that higher prediction scores typically align with increased feature magnitudes, an effect that is more pronounced in stress-inducing tasks compared to neutral ones. Additionally, the model occasionally highlights short segments in the neutral tasks as seen in the first row in Figure 2.

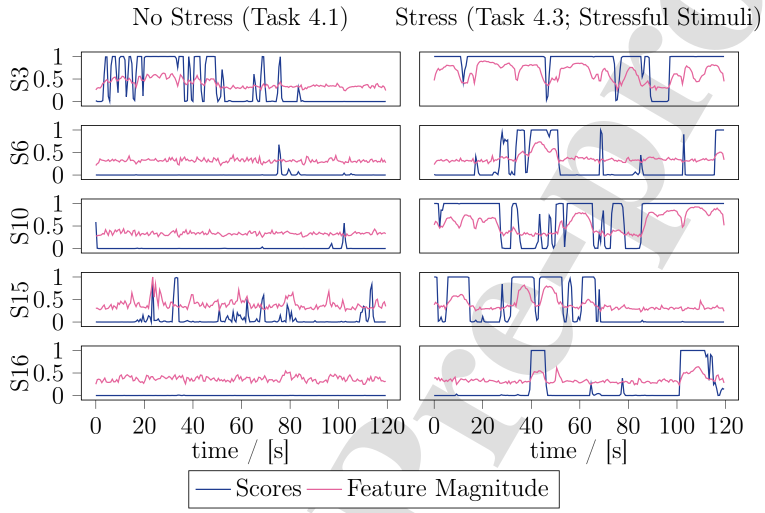


Figure 2: Scores and feature magnitudes of the snippet classification head during the neutral task 4.1 (left) and stressful stimuli task 4.3 (right). A time interval of 2 minutes is shown for the outputs of five sample subjects. Modified figure from [20]

Correlations between predictions and action unit time series To better understand which facial movements the network focuses on, we calculated the correlation between network prediction scores and 17 facial Action Unit (AU) time series for each subject and task in our stress dataset, identifying the action units that were most pronounced in the time steps with the highest network scores. All correlation results were corrected for false detection rate and are visualised as box plots in the Supplementary Material in Fig. 2. The findings show that the AUs correlated with network predictions vary across different task combinations. However, AU 14 (Dimpler, see Supplementary Material, Fig. 3d)) stands out as consistently prominent across most tasks. Except for task 4.2 vs. 4.3, more than 10 participants showed at least a moderate correlation of 40 % or higher between AU 14 and the prediction scores. In the neuropsychological test tasks (Mental Workload, tasks 3.1 vs. 3.2 and 3.1 vs. 3.3), AU 10 and AU 12 (Upper Lip Raiser and Lip Corner Puller, see Supplementary Material, Fig. 3b) and 3c)) were also notably correlated with model predictions, suggesting their relevance for stress-related expressions in cognitively demanding tasks. AU 4 (Brow Lowerer, Supplementary Material, Fig. 3a)) is particularly present in many test subjects in all tasks except 1.2 vs. 1.3 and 2.1 vs. 2.2, indicating its general relevance for stress detection in facial expressions.

7. Discussion

7.1. Discussion of the Results

Our experiments demonstrate that the proposed MIL-based approach yields robust performance across diverse stress detection tasks, particularly when combined with the MSMHN architecture and bottom-k sampling. The use of bottom-k features, i.e., snippets from stress videos with the lowest predicted stress scores, proved particularly beneficial in our stress dataset. These snippets likely represent neutral or low-stress moments within stressful tasks and increase dataset diversity by introducing intra-task variation. Variations in head posture, facial orientation, or brief relaxation phases may help the model disentangle task-specific but non-stress-related factors from genuine stress expressions, thereby reducing overfitting to incidental cues. In contrast, dense snippet-level supervision did not yield reliable stress classification, supporting the assumption that stress is not consistently visible at every time point in a video, further motivating the use of MIL.

We further observe that conditioning the network on speech significantly improves classification performance in both datasets. In our stress dataset, where stress and neutral tasks are evenly distributed within both passive and active task groups, the large improvement cannot be attributed to class imbalance and instead underscores the value of contextual cues that serve to normalise facial dynamics. However, in the StressID dataset, part of the improvement may stem from conditioning acting as a strong prior, as only a single passive task is labelled as stressful. Nevertheless, conditioning also resulted in performance gains in the self-assessment label setting, suggesting that also in this dataset, approximate contextual information can disambiguate subtle or noisy stress signals. Across all tested conditioning variants, pre- and post-temporal conditioning lead to comparable performance, suggesting that conditioning primarily mitigates task-induced variability rather than reshaping higher-level temporal representations. Explicit task conditioning does not provide substantial gains and is a modelling aid only under controlled data collection protocols. Consequently, we rely on speech-based conditioning alone, which is both effective and practical, as it could be robustly inferred using standard speech activity detection or lightweight speech classifiers in real-world settings. Finally, because we divided the StressID videos into 20 s-sub-segments, the classification accuracy could potentially be further improved by aggregating predictions, e.g., through majority voting or confidence-weighted averaging.

7.2. Comparison with State-of-the-Art methods

Recent stress detection methods differ in their assumptions and supervision. For instance, Giannakakis et al. [7] achieve strong performance using a pairwise transformation that requires subject-specific baseline recordings, achieving 84.7 % accuracy without and 88.6 % with this normalisation in the *all tasks* setting. While highly effective, particularly for *single-task* settings, where accuracies reach up to 99.19 %, it limits applicability when such reference data is unavailable. In contrast, our approach does not rely on subject-level baselines and achieves higher accuracy without pairwise normalisation (96.9 %) in the *all tasks* setting, while remaining competitive across individual task categories. On the StressID dataset, our video-only model performs comparably to the unimodal video baseline in [21] (F1: 69 % vs. 70 %, ACC: 70 %). Although multimodal fusion improves results in [21], our focus in this work is deliberately on video-only stress modelling; incorporating additional modalities remains a promising direction for future work.

7.3. Limitations

While the results demonstrate strong stress detection capabilities across diverse scenarios, some limitations and open research questions remain.

Lack of snippet-level evaluation Without fine-grained snippet-level labels, it cannot be guaranteed that the model only attends to facial cues directly indicative of stress. This becomes evident in the mental load task, where participants tend to smile after making errors. Although smiling is not itself a stress marker, the model may learn to associate it with stressful situations due to its temporal proximity to task-related errors. This is supported by our correlation analysis (Supplementary Material, Fig. 2), which revealed stronger correlations between the network predictions and AUs 6, 10, and 12 (cheek raiser, upper lip raiser, and lip corner puller, see Supplementary Material, Fig. 3) in neuropsychological tasks compared to in other task combinations. These action units are commonly activated during smiling, highlighting the risk of the model focusing on indirect cues.

Hyperparameter k The choice of the hyperparameter k is inherently dataset-dependent [57]. Consequently, a value of k that is optimal for one dataset may not transfer well to others and should be selected with consideration of the expected frequency of anomalies. If k is too low, the model may focus on too few examples; if too high, irrelevant or contradictory snippets may be included. To assess the sensitivity of our method to this hyperparameter, we performed an ablation study over different values of k (Figure 2 in the Supplementary Material). The results show that model performance is largely stable across different choices of k and does not degrade substantially even when the selected top and bottom snippets partially overlap. However, we note that our approach is primarily designed for the stress elicitation scenarios considered in this work, rather than uninterrupted peak stress, as it assumes the presence of snippets less indicative of stress.

Snippet feature extraction In this work, we do not explicitly optimise the snippet feature extraction and instead rely on a pre-trained contrastive learning model to obtain snippet representations. While this choice is well aligned with our MIL setting, there could be other models that further improve classification. Common approaches for video representation learning include 3D convolutional networks such as C3D and I3D [53, 54], two-stream architectures combining appearance and motion cues [64], and transformer-based models capturing long-range temporal dependencies [**?** 65]. However, compared to standard video encoders such as I3D, the contrastive model used has shown superior performance on facial video data in related work [55]. An alternative direction would be to use pretrained large-scale video foundation models, such as VideoMAE [66] or Video Swin Transformers [67], which may capture complementary spatiotemporal dynamics but typically require substantially larger datasets or task-specific fine-tuning. In addition, since facial landmarks are treated as temporal signals rather than raw visual inputs, our approach is related to general pre-training for time-series representation learning. Well-established signal-based pretraining strategies include contrastive predictive coding [68], temporal contrastive learning [69], masked modelling for time series [70], and more recent temporal memory fusion approaches that explicitly model long-term dependencies [71]. Such methods could be naturally extended to landmark-based facial analysis and offer an interesting direction for future work.

StressID self-assessment labels in MIL For the StressID dataset, we observed a slightly lower performance with self-assessment labels compared to task labels. This could be explained by the subjective nature of stress perception. Even if a participant reports low overall stress for a task, there may be short episodes of stress behaviour within the video. By focusing on the top-scoring snippets, the MIL framework could capture these short high-stress segments, possibly conflicting with the global label. Especially for self-assessment labels, this mismatch introduces ambiguity.

Dataset recording conditions Finally, it should be noted that all data were collected in a controlled lab environment. While this ensures consistency and high-quality facial landmarks, it may limit generalisability to real-world conditions. However, the use of contrastive learning for feature extraction, especially on landmark data, reduces sensitivity to variations in appearance, lighting, and background.

Interpretability Finally, relying on landmark data results in primarily temporal interpretability, which we complement with a correlation analysis to link temporal representations to facial Action Units. We acknowledge that spatial interpretability at the level of facial regions remains an important direction for future work.

8. Conclusion

We proposed a video-based stress detection framework built on top-k MIL, a temporal feature extractor with multi-head attention, and a task-type conditioning mechanism. The approach assumes that stress-related behaviour may appear only in short segments of a video, even within tasks labelled as stressful, and improves robustness by incorporating both top-k and bottom-k snippets. Our ablation study systematically evaluated the contributions of MIL, the temporal feature network, the integration of bottom-k snippets, and the conditioning module. With the full set of proposed components, our model achieved an average accuracy of 95.46% and F1 score of 95.49% on our stress dataset. When trained on all tasks jointly, accuracy and F1 dropped to 62.51 % and 63.09%, respectively, but were recovered to 96.88% and 96.75% using FiLM conditioning. On the StressID dataset, our method achieved 78.16% accuracy and 78.12 % F1 score using task labels and 67.92 % accuracy and 68.50 % F1 score using self-assessments. Linear conditioning improved performance to 81.39 % accuracy and 83.72 % (task labels), and 71.29 % accuracy and 71.95 % F1 score (self-assessments).

Beyond performance, MIL enables interpretable stress localisation by identifying video snippets most relevant for classification. Our analyses revealed correlations between network predictions and specific facial action units, offering insights into the facial dynamics of stress.

CRediT authorship contribution statement

Nele Sophie Brügge: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Software, Visualization, Writing - original draft. **Alexandra Korda**: Validation, Writing - Review & Editing **Heinz Handels**: Funding acquisition, Supervision, Writing - Review & Editing **Giorgos Giannakakis**: Data Curation, Project administration, Writing - Review & Editing

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Funding

Research activities leading to this publication have been financially supported by the Bundesministerium Forschung, Technologie und Raumfahrt, Germany (BMFTR) (16KISA050K).

Ethics approval

The Research Ethics Committee of FORTH provided its approval for this study (approval no. 155/12-09-2022).

Consent

Each participant gave their free and informed consent.

References

[1] J. E. Dimsdale, Psychological stress and cardiovascular disease, Journal of the American College of Cardiology 51 (13) (2008) 1237–1246. doi:10.1016/j.jacc.2007.12.024.
URL https://www.sciencedirect.com/science/article/pii/S0735109708002581

[2] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, IEEE Transactions on Affective Computing 13 (1) (2019) 440–460.

[3] G. Giannakakis, M. Pediaditis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, M. Tsiknakis, Stress and anxiety detection using facial cues from videos, Biomed. Signal Process. Control. 31 (2017) 89–101.

[4] A. I. Korda, G. Giannakakis, E. Ventouras, P. A. Asvestas, N. Smyrnis, K. Marias, G. K. Matsopoulos, Recognition of blinks activity patterns during stress conditions using cnn and markovian analysis, Signals 2 (1) (2021) 55–71.

[5] F. Bevilacqua, H. Engström, P. Backlund, Automated analysis of facial cues from videos as a potential method for differentiating stress and boredom of players in games, International Journal of Computer Games Technology 2018 (2018).

[6] C. Daudelin-Peltier, H. Forget, C. Blais, A. Deschênes, D. Fiset, The effect of acute social stress on the recognition of facial expression of emotions, Scientific Reports 7 (1) (2017) 1036.

[7] G. Giannakakis, A. Roussos, C. Andreou, S. Borgwardt, A. I. Korda, Stress recognition identifying relevant facial action units through explainable artificial intelligence and machine learning, Computer Methods and Programs in Biomedicine 259 (2025) 108507. doi:10.1016/j.cmpb.2024.108507.
URL https://www.sciencedirect.com/science/article/pii/S0169260724005005

[8] W. Sultani, C. Chen, M. Shah, Real-world anomaly detection in surveillance videos, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, Salt Lake City, UT, 2018, pp. 6479–6488. doi:10.1109/CVPR.2018.00678.
URL https://ieeexplore.ieee.org/document/8578776/

[9] J. Zhang, L. Qing, J. Miao, Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection, in:

2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 4030–4034. doi:10.1109/ICIP.2019.8803657.

[10] B. Wan, Y. Fang, X. Xia, J. Mei, Weakly supervised video anomaly detection via center-guided discriminative learning, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1–6. doi:10.1109/ICME46284.2020.9102722.

[11] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans, G. Carneiro, Weakly-supervised video anomaly detection with robust temporal feature magnitude learning, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, Montreal, QC, Canada, 2021, pp. 4955–4966. doi:10.1109/ICCV48922.2021.00493.
URL https://ieeexplore.ieee.org/document/9710957/

[12] J.-C. Feng, F.-T. Hong, W.-S. Zheng, MIST: Multiple instance self-training framework for video anomaly detection, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, 2021, pp. 14004–14013. doi:10.1109/CVPR46437.2021.01379.
URL https://ieeexplore.ieee.org/document/9578773/

[13] S. Li, F. Liu, L. Jiao, Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 1395–1403. doi:10.1609/aaai.v36i2.20028.
URL https://ojs.aaai.org/index.php/AAAI/article/view/20028

[14] W. Luo, W. Liu, S. Gao, A revisit of sparse coding based anomaly detection in stacked RNN framework, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, 2017, pp. 341–349. doi:10.1109/ICCV.2017.45.
URL http://ieeexplore.ieee.org/document/8237307/

[15] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, G. Li, Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 1237–1246. doi:10.1109/CVPR.2019.00133.
URL https://ieeexplore.ieee.org/document/8953791/

[16] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, Z. Yang, Not only look, but also listen: Learning multimodal violence detection under weak supervision, in: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Springer-Verlag, Berlin, Heidelberg, 2020, pp. 322–339. doi:10.1007/978-3-030-58577-8_20.
URL 10.1007/978-3-030-58577-8_20

[17] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 1975–1981, ISSN: 1063-6919. doi:10.1109/CVPR.2010.5539872.

[18] J. J. Gross and R. W. Levenson, "Hiding feelings: The acute effects of inhibiting negative and positive emotion," *Journal of Abnormal Psychology*, vol. 106, no. 1, pp. 95–103, 1997.

[19] W. Li, N. Vasconcelos, Multiple instance learning for soft bags via top instances, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 4277–4285. doi:10.1109/CVPR.2015.7299056.

[20] N. S. Brügge, A. I. Korda, S. J. Borgwardt, C. Andreou, G. A. Giannakakis, H. Handels, Bag-level multiple instance learning for acute stress detection from video data, in: J. Kim, R. C. Conceição, M. Yousef, A. Bhavsar, S. Pelayo, A. Fred, H. Gamboa (Eds.), Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2025 - Volume 2: HEALTHINF, Porto, Portugal, February 20-22, 2025, SCITEPRESS, 2025, pp. 285–296. doi:10.5220/0013364900003911.
URL 10.5220/0013364900003911

[21] H. Chaptoukaev, V. Strizhkova, M. Panariello, B. Dalpaos, A. Reka, V. Manera, S. Thümmler, E. ISMAILOVA, N. W., f. bremond, M. Todisco, M. A. Zuluaga, L. M. Ferrari, Stressid: a multimodal dataset for stress identification, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Vol. 36, Curran Associates, Inc., New Orleans, Louisiana, USA, 2023, pp. 29798–29811.

[22] M. Naegelin, R. P. Weibel, J. I. Kerr, V. R. Schinazi, R. La Marca, F. von Wangenheim, C. Hoelscher, A. Ferrario, An interpretable machine learning approach to multimodal stress detection in a simulated office environment, Journal of Biomedical Informatics 139 (2023) 104299. doi:10.1016/j.jbi.2023.104299.
URL https://www.sciencedirect.com/science/article/pii/S1532046423000205

[23] P. Bobade, M. Vani, Stress detection with machine learning and deep learning using multimodal physiological data, in: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), 2020, pp. 51–57. doi:10.1109/ICIRCA48905.2020.9183244.
URL https://ieeexplore.ieee.org/abstract/document/9183244

[24] A. I. Siam, S. A. Gamel, F. M. Talaat, Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques, Neural Computing and Applications 35 (17) (2023) 12891–12904. doi:10.1007/s00521-023-08428-w.
URL 10.1007/s00521-023-08428-w

[25] P. Garg, J. Santhosh, A. Dengel, S. Ishimaru, Stress detection by machine learning and wearable sensors, in: 26th International Conference on Intelligent User Interfaces - Companion, IUI '21 Companion, Association for Computing Machinery, New York, NY, USA, 2021, pp. 43–45. doi:10.1145/3397482.3450732.
URL https://dl.acm.org/doi/10.1145/3397482.3450732

[26] S. Hosseini, S. Katragadda, R. T. Bhupatiraju, Z. Ashkar, C. Borst, K. Cochran, R. Gottumukkala, A multi-modal sensor dataset for continuous stress detection of nurses in a hospital (2021). doi:10.5061/dryad.5hqbzkh6f.
URL https://zenodo.org/record/5514277

[27] C. Viegas, S.-H. Lau, R. Maxion, A. Hauptmann, Towards independent stress detection: A dependent model using facial action units, 2018 International Conference on Content-Based Multimedia Indexing (CBMI) (2018) 1–6doi:10.1109/CBMI.2018.8516497.
URL https://ieeexplore.ieee.org/document/8516497/

[28] R. Li, Z. Liu, Stress detection using deep neural networks, BMC Medical Informatics and Decision Making 20 (2020). doi:10.1186/s12911-020-01299-4.

[29] J. Zhang, H. Yin, J. Zhang, G. Yang, J. Qin, L. He, Real-time mental stress detection using multimodality expressions with a deep learning framework, Frontiers in Neuroscience 16 (2022) 947168. doi:10.3389/fnins.2022.947168.

[30] H. Zhang, L. Feng, N. Li, Z. Jin, L. Cao, Video-based stress detection through deep learning, Sensors 20 (19) (2020) 5552. doi:10.3390/s20195552.

[31] S. Kumar, A. S. M. Iftekhar, M. Goebel, T. Bullock, M. Maclean, M. Miller, T. Santander, B. Giesbrecht, S. Grafton, B. Manjunath, Stressnet: Detecting stress in thermal videos, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, pp. 998–1008. doi:10.1109/WACV48630.2021.00104.

[32] T. Jeon, H. B. Bae, Y. Lee, S. Jang, S. Lee, Deep-learning-based stress recognition with spatial-temporal facial information, Sensors 21 (22) (2021) 7498. doi:10.3390/s21227498.

[33] B. Hasani, M. H. Mahoor, Facial expression recognition using enhanced deep 3d convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 30–40.

[34] M. Gavrilescu, N. Vizireanu, Predicting depression, anxiety, and stress levels from videos using the facial action coding system, Sensors 19 (17) (2019) 3693, publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/s19173693.
URL https://www.mdpi.com/1424-8220/19/17/3693

[35] G. Giannakakis, M. R. Koujan, A. Roussos, K. Marias, Automatic stress detection evaluating models of facial action units, in: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, Buenos Aires, Argentina, 2020, pp. 728–733. doi:10.1109/FG47880.2020.00129.
URL https://ieeexplore.ieee.org/document/9320268/

[36] E. Turcan, S. Muresan, K. McKeown, Emotion-infused models for explainable psychological stress detection, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2895–2909. doi:10.18653/v1/2021.naacl-main.230.
URL https://aclanthology.org/2021.naacl-main.230

[37] T. Tong, R. Wolz, Q. Gao, J. Hajnal, D. Rueckert, Multiple instance learning for classification of dementia in brain MRI, Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention 16 (2013) 599–606. doi:10.1016/j.media.2014.04.006.

[38] M. Kandemir, F. A. Hamprecht, Computer-aided diagnosis from weak supervision: a benchmarking study, Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society 42 (2015) 44–50. doi:10.1016/j.compmedimag.2014.11.010.

[39] S. Geng, S. Jia, Y. Qiao, J. Yang, Z. Jia, Combining CNN and MIL to assist hotspot segmentation in bone scintigraphy, in: S. Arik, T. Huang, W. K. Lai, Q. Liu (Eds.), Neural Information Processing, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2015, pp. 445–452. doi:10.1007/978-3-319-26561-2_53.

[40] H. Li, F. Yang, X. Xing, Y. Zhao, J. Zhang, Y. Liu, M. Han, J. Huang, L. Wang, J. Yao, Multi-modal multi-instance learning using weakly correlated histopathological images and tabular clinical information, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 529–539.

[41] M. Kandemir, C. Zhang, F. A. Hamprecht, Empowering multiple instance histopathology cancer diagnosis by cell graphs, in: P. Golland, N. Hata, C. Barillot, J. Hornegger, R. Howe (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014, pp. 228–235. doi:10.1007/978-3-319-10470-6_29.

[42] M. M. Dundar, S. Badve, V. C. Raykar, R. K. Jain, O. Sertel, M. N. Gurcan, A multiple instance learning approach toward optimal clas-

sification of pathology slides, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 2732–2735, ISSN: 1051-4651. doi:10.1109/ICPR.2010.669.
URL https://ieeexplore.ieee.org/document/5596023

[43] Y. Xu, J.-Y. Zhu, E. Chang, Z. Tu, Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 964–971, ISSN: 1063-6919. doi:10.1109/CVPR.2012.6247772.
URL https://ieeexplore.ieee.org/document/6247772

[44] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, E. I.-C. Chang, Deep learning of feature representation with multiple instance learning for medical image analysis, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 1626–1630, ISSN: 2379-190X. doi:10.1109/ICASSP.2014.6853873.

[45] Y. Xu, J.-Y. Zhu, E. Chang, M. Lai, Z. Tu, Weakly supervised histopathology cancer image segmentation and classification, Medical image analysis 18 (2014) 591–604. doi:10.1016/j.media.2014.01.010.

[46] K. Sikka, A. Dhall, M. S. Bartlett, Classification and weakly supervised pain localization using multiple segment representation, Image and Vision Computing 32 (10) (2014) 659–670. doi:10.1016/j.imavis.2014.02.008.
URL https://www.sciencedirect.com/science/article/pii/S0262885614000456

[47] K. Sikka, A. Dhall, M. Bartlett, Weakly supervised pain localization using multiple instance learning, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013, pp. 1–8. doi:10.1109/FG.2013.6553762.

[48] Y. Wang, J. Ma, B. Hao, P. Hu, X. Wang, J. Mei, S. Li, Automatic depression detection via facial expressions using multiple instance learning, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1933–1936. doi:10.1109/ISBI45749.2020.9098396.
URL https://ieeexplore.ieee.org/document/9098396

[49] Y. Tian, G. Pang, F. Liu, Y. Liu, C. Wang, Y. Chen, J. Verjans, G. Carneiro, Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 88–98. doi:10.1007/978-3-031-16437-8_9.
URL 10.1007/978-3-031-16437-8_9

[50] github, Github (2020).
URL https://github.com/ggian/stress_dataset

[51] J. R. Stroop, Studies of interference in serial verbal reactions., Journal of experimental psychology 18 (6) (1935) 643.

[52] D. Gronwall, Paced auditory serial-addition task: a measure of recovery from concussion, Perceptual and motor skills 44 (2) (1977) 367–373.

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Santiago, Chile, 2015, pp. 4489–4497. doi:10.1109/ICCV.2015.510.
URL http://ieeexplore.ieee.org/document/7410867/

[54] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA, 2017, pp. 4724–4733. doi:10.1109/CVPR.2017.502.
URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.502

[55] N. S. Brügge, E. Mohammadi, A. Münchau, T. Bäumer, C. Frings, C. Beste, V. Roessner, H. Handels, Towards privacy and utility in tourette TIC detection through pretraining based on publicly available video data of healthy subjects, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023. doi:10.1109/ICASSP49357.2023.10095309.

[56] B. Angles, Y. Jin, S. Kornblith, A. Tagliasacchi, K. M. Yi, MIST: Multiple instance spatial transformer, in: 2021 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), IEEE, Nashville, TN, USA, 2021, pp. 2412–2422. doi:10.1109/CVPR46437.2021.00244. URL https://ieeexplore.ieee.org/document/9578563/

[57] W. Li, N. Vasconcelos, Multiple instance learning for soft bags via top instances, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, 2015, pp. 4277–4285. doi:10.1109/CVPR.2015.7299056.
URL http://ieeexplore.ieee.org/document/7299056/

[58] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. on Artificial Intelligence*, 2018.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.

[60] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2018. doi:10.1109/CVPR.2018.00813.

[61] Y. Yang, Y. Su, S. An, VDSSA: Ventral & Dorsal Sequential Self-attention AutoEncoder for Cognitive-Consistency Disentanglement, in: Pattern Recognition and Computer Vision, Springer, 2022, pp. 693–705. doi:10.1007/978-3-031-18910-4_55.

[62] Q. Peng, S. Zhu, Y. Su, M. Xing, Gaze-and-Machine Dual-Driven Attention Fusion Network for Medical Image Classification, in: Advanced Intelligent Computing Technology and Applications, Springer, 2025, pp. 402–412. doi:10.1007/978-981-95-0036-9_34.

[63] K. Hara, H. Kataoka, Y. Satoh, Learning spatio-temporal features with 3d residual networks for action recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3154–3160.

[64] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proceedings of the 28th International Con-

ference on Neural Information Processing Systems (NIPS), MIT Press, 2014, pp. 568–576.

[65] G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding?, in: Proceedings of the International Conference on Machine Learning (ICML), 2021.

[66] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, in: Advances in Neural Information Processing Systems (NeurIPS), 2022.

[67] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video Swin Transformer, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2022, pp. 3192–3201. doi:10.1109/CVPR52688.2022.00320.

[68] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748, 2018.
URL https://arxiv.org/abs/1807.03748

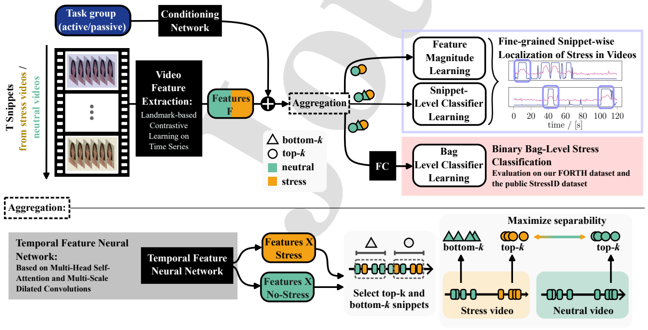
[69] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. K. Kwoh, X. Li, C. Guan, Time-series representation learning via temporal and contextual contrasting, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), 2021, pp. 2352–2359.
URL https://www.ijcai.org/proceedings/2021/324

[70] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, C. Eickhoff, A transformer-based framework for multivariate time series representation learning, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), ACM, 2021, pp. 2114–2124. doi:10.1145/3447548.3467401.
URL https://doi.org/10.1145/3447548.3467401

[71] Y. Wang, Y. Zhao, General pre-trained inertial signal feature extraction based on temporal memory fusion, Information Fusion, 2025.
URL https://www.sciencedirect.com/science/article/pii/S1566253525003471

1 HIGHLIGHTS

- Video-based facial cues enable non-invasive detection of acute stress
- Learning from weak labels identifies brief stress moments in long videos
- Combining top- and bottom-k video segments improves stress classification
- Attention modelling captures short- and long-term facial stress patterns
- Conditioning on speech activity improves stress detection accuracy


Task group (active/passive)
Conditioning Network
T Snippets
from stress videos /
neutral videos
Video Feature Extraction:
Landmark-based Contrastive Learning on Time Series
Features F
Aggregation
△ bottom-k
○ top-k
🟩 neutral
🟧 stress
Feature Magnitude Learning
Snippet-Level Classifier Learning
Fine-grained Snippet-wise Localization of Stress in Videos
0 20 40 60 80 100 120
time / [s]
FC
Bag Level Classifier Learning
Binary Bag-Level Stress Classification
Evaluation on our FORTH dataset and the public StressID dataset
Aggregation:
Temporal Feature Neural Network:
Based on Multi-Head Self-Attention and Multi-Scale Dilated Convolutions
Temporal Feature Neural Network
Features X Stress
Features X No-Stress
Select top-k and bottom-k snippets
Maximize separability
bottom-k
top-k
top-k
Stress video
Neutral video

CRediT authorship contribution statement

Nele Sophie Brügge: Conceptualization, Methodology, Investigation, Validation, Formal analysis, Software, Visualization, Writing - original draft.

Alexandra Korda: Validation, Writing - Review & Editing

Heinz Handels: Funding acquisition, Supervision, Writing - Review & Editing

Giorgos Giannakakis: Data Curation, Project administration, Writing - Review & Editing

Funding

Research activities leading to this publication have been financially supported by the Bundesministerium Forschung, Technologie und Raumfahrt, Germany (BMFTR) (16KISA050K).

Acknowledgements

None

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: