







Bag-Level Multiple Instance Learning for Acute Stress Detection from Video Data

Nele Sophie Brügge¹^a, Alexandra Korda²^b, Stefan Borgwardt²^c, Christina Andreou²^d,
Giorgos Giannakakis^{3,4,5}^e and Heinz Handels^{1,6}^f

¹German Research Center for Artificial Intelligence, AI in Medical Image and Signal Processing, Lübeck, Germany

²Translational Psychiatry, Department of Psychiatry and Psychotherapy, University of Luebeck, Lübeck, 23562, Germany

³Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), Heraklion, Greece

⁴Department of Electronic Engineering, Hellenic Mediterranean University, Chania, Greece

⁵Institute of Agri-food and Life Sciences, University Research and Innovation Center, Hellenic Mediterranean University, Heraklion, Greece

⁶Institute of Medical Informatics, University of Luebeck, Lübeck, Germany

Keywords: Stress Detection, Multiple Instance Learning, Video Analysis, Neural Networks, Machine Learning.

Abstract: Stress detection is a complex challenge with implications for health and well-being. It often relies on sensors recording biomarkers and biosignals, which can be uncomfortable and alter behaviour. Video-based facial feature analysis offers a noninvasive alternative. This study explores video-level stress detection using top- k Multiple Instance Learning applied to medical videos. The approach is motivated by the assumption that subjects partly show normal behaviour while performing stressful experimental tasks. Our contributions include a tailored temporal feature network and optimised data utilisation by additionally incorporating bottom- k snippets. Leave-five-subjects-out stress detection results of 95.46 % accuracy and 95.49 % F1 score demonstrate the potential of our approach, outperforming the baseline methods. Additionally, through multiple instance learning, it is possible to show which temporal video segments the network pays particular attention to.

1 INTRODUCTION


Stress is a psychological response to overdemanding events that are perceived as threatening or challenging. It can have negative effects on one's physical and mental health. Recognition of stress is challenging and commonly based on the evaluation of a variety of biomarkers (i.e. cortisol, corticotropin-releasing factor (CRF), and adrenocorticotropin (ACTH) (Chrousos, 2009)) and biosignals (features derived from ECG, EDA, respiration, EMG, etc.) (Giannakakis et al., 2019). However, the evaluation of biosignals and biomarkers requires the use of sensors, which can be invasive and uncomfortable


and may alter the subject's response to stress.


In recent years, there has been increasing interest in detecting stress based on facial features, which in most cases does not meet the performance of that including biosignals. Yet, video monitoring of the subjects represents a convenient and noninvasive alternative. Besides, for a more objective facial stress recognition, there has been an effort for the identification of involuntary or semi-voluntary facial parameters (Giannakakis et al., 2017; Korda et al., 2021; Bevilacqua et al., 2018; Daudelin-Peltier et al., 2017), (Giannakakis et al., 2025). These include blinks, mouth micro activity or micro-expressions.


Still, it is not yet fully understood how different types of stress are manifested in facial expressions and the expression of stress can vary greatly between individuals in terms of intensity and type. The detection of stress is therefore one of many medical tasks for which it is challenging to create fine-grained labelled datasets. Furthermore, labelling would have to


^a <https://orcid.org/0009-0006-2039-423X>

^b <https://orcid.org/0000-0001-8843-4951>

^c <https://orcid.org/0000-0002-5792-3987>

^d <https://orcid.org/0000-0002-6656-9043>

^e <https://orcid.org/0000-0002-0958-5346>

^f <https://orcid.org/0000-0002-3499-4328>

be done by experts, is time consuming and cumbersome, especially for video data.

To address these challenges, this work proposes the use of Multiple Instance Learning (MIL) for the application to stress detection from video data. Its main advantages are that it requires only video-level labels and that it can also detect subtle, short-term anomalies in longer videos. Videos that contain a target event are labelled as positive, while other videos are labelled as negative. The assumption in MIL is that videos labelled as positive also contain negative segments, while videos labelled as negative consist only of negative instances. MIL is typically used to detect anomalies in surveillance camera videos (Sultani et al., 2018), (Zhang et al., 2019), (Wan et al., 2020), (Tian et al., 2021), (Feng et al., 2021), (Li et al., 2022) (ShanghaiTech (Luo et al., 2017), (Zhong et al., 2019), UCF-Crime (Sultani et al., 2018), XD-Violence (Wu et al., 2020) and UCSD-Peds (Mahadevan et al., 2010)).

Regarding stress detection, we consider MIL an appropriate method considering that participants' faces remain neutral for many frames even in stressful tasks, presenting only short periods of stressful facial behaviour. MIL can further be utilized to provide not only video-level but also snippet-level (set of few frames) predictions, providing explainability and insights into temporal dynamics of stress behaviour. Our approach is motivated by top- k MIL (Li and Vasconcelos, 2015a), (Tian et al., 2021), which trains a classifier using the k instances with the highest classification scores as positive instances. The trained model can be used to classify new snippets into labels (stress, no stress) based on the features that provide the most representative snippet instances.

We make modifications to top- k MIL, including the use of an appropriate feature extraction method, the use of bottom- k snippets for MIL, and the design of a tailored temporal attention network and a bag-level classification network for the binary classification task. As feature extraction method, we use a contrastive learning network pretrained on facial landmarks from video data. The input videos are cut and divided into negative (no stress) and positive (stress presence) bags for the second MIL training phase. We propose an attention-based network for temporal feature extraction, that captures long- and short-term facial expression patterns. Our training scheme also includes the bottom- k snippets of positive bags by assigning them the neutral label to make the best use of the limited available data and to improve the robustness of our model. This is based on the assumptions that there are phases of neutral behaviour also in videos showing subjects during stressful experimen-

tal tasks and that the snippets with the lowest feature norms most likely represent neutral snippets. In summary, our contributions consist of

- Applying MIL to stress detection from video data
- Proposing an MIL approach that exploits both top- k and bottom- k video snippets in training
- Designing a temporal feature extraction network with multi-head attention.

2 RELATED WORK

2.1 Stress Detection Using Machine Learning

Recent research in stress detection using machine learning has explored a spectrum of methods. Conventional ML approaches, such as Random Forests and Support Vector Machines, have been employed effectively (Naegelin et al., 2023), (Bobade and Vani, 2020), (Siam et al., 2023), (Garg et al., 2021), (Hosseini et al., 2021), (Viegas et al., 2018). These studies used data from a variety of sensors, including wearables, electrodermal activity, electrocardiography, electroencephalography and temperature (Li and Liu, 2020), (Naegelin et al., 2023), (Bobade and Vani, 2020), (Siam et al., 2023), (Hosseini et al., 2021), (Garg et al., 2021), (Zhang et al., 2022). In parallel, video data analysis (Zhang et al., 2022), (Zhang et al., 2020), (Kumar et al., 2021), (Jeon et al., 2021) has emerged as a convenient and non-invasive alternative for stress detection, providing a good reproducibility without the requiring a precise sensor placement.

The analysis of video data has greatly benefited from advancements in complex neural network architectures, achieving high accuracy in facial stress recognition (Hasani and Mahoor, 2017), (Jeon et al., 2021), (Kumar et al., 2021), (Li and Liu, 2020), (Zhang et al., 2020), (Zhang et al., 2022). As an example, in (Hasani and Mahoor, 2017), a 3D Convolutional Neural Network method for facial expression recognition in videos was proposed, yielding a stress recognition accuracy up to 90%. Using also information from voice and ECG of 20 participants, in (Zhang et al., 2022) a neural network based on I3D features and a temporal attention module was proposed, achieving an accuracy of 85.1%. Other 2D ResNet-based approaches use temporal attention (Jeon et al., 2021) or long short-term memory (LSTM) layers (Zhang et al., 2020), (Kumar et al., 2021) to introduce temporal information. Instead of applying neural networks directly to the raw video data, some stud-

ies have focused on extracting facial action units from videos as input features for classification (Gavrilescu and Vizireanu, 2019), (Giannakakis et al., 2020).

Additionally, the stress detection task has been investigated across diverse environments and scenarios, spanning office settings (Naegelin et al., 2023), hospital scenarios (Hosseini et al., 2021), activities like car driving (Siam et al., 2023) or social media posts (Turcan et al., 2021). In previous work, stressful tasks often consist of a mental task, memory task, arithmetic task, or external stimuli such as noisy sounds, showing arousing photos or videos and physical stimuli. While many approaches to stress detection have been extensively investigated, the potential of Multiple Instance Learning (MIL) remains largely unexplored. In this paper we evaluate MIL on six different tasks and stimuli. Further, the use of multiple instance learning in stress detection still remains largely unexplored, although many approaches have been extensively investigated.

2.2 Multiple Instance Learning in Medical Image and Video Analysis

MIL has shown promising results in many medical image and video analysis applications. Examples in medical image analysis include dementia classification in brain MRI (Tong et al., 2013), diabetic retinopathy detection in colour fundus images (Kandemir and Hamprecht, 2015) and hotspot detection in bone scintigraphy images (Geng et al., 2015). Several studies have applied MIL to histopathology patches in cancer research, for example to detect lymph node metastases in breast cancer (Li et al., 2021), (Kandemir et al., 2014), (Dundar et al., 2010) and the classification of esophagus (Kandemir and Hamprecht, 2015), (Kandemir et al., 2014) or colon cancer (Xu et al., 2012), (Xu et al., 2014a), (Xu et al., 2014b).

There is also work on medical video analysis, while MIL is more commonly applied to anomaly detection in surveillance camera videos. Sikka et al. (Sikka et al., 2014), (Sikka et al., 2013) used a weakly supervised MIL approach for pain localisation from medical videos. In (Wang et al., 2020), MIL was used to detect depression from videos using facial landmarks. Further, (Tian et al., 2022) proposed a contrastive transformer-based approach for weakly supervised polyp frame detection in colonoscopy videos. To the best of our knowledge, MIL has not been applied to detect stress from facial video.

3 STRESS DATASET

We recorded videos of subjects performing different stressful tasks. The experimental protocol was designed to investigate facial and physiological responses under stress conditions. The experimental dataset comprised 58 individuals (24 men and 34 women) with an average age of 26.9 ± 4.8 years.

3.1 Video Acquisition Protocol

All participants were seated in front of a monitor and a camera. The camera's field of view covered the participant's face. Possible movements during the experiment were taken into account. The camera was mounted on a tripod and positioned at the back of the screen at a distance of about 90 cm from the face. Ambient lighting conditions were ensured to reduce the effects of specular lighting. The videos had a sampling rate of 60 frames per second and a resolution of 1216 x 1600 pixels, which were subsampled to 608 x 800 pixels at 30 frames per second.

3.2 Experimental Tasks

The experiment included neutral tasks (used as reference) and stressful tasks in which stress conditions were simulated and induced using different types of stressors. These stressors were categorised into 4 different 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. Each participant completed eleven tasks: four in neutral, six in stressed, and one in a relaxed state. Every experiment began with a neutral or relaxing phase at each stage as baseline and each recording had a duration of 2 min.

The social exposure phase included an interview asking the participant to describe him/herself. It originated from the stress of exposure that an actor faces when she/he is on stage. The reference for this phase was the participant saying conventional words (e.g. counting from one to ten, listing the months of the year, etc.). The emotional recall phase included stress elicitation by asking participants to recall and relive a stressful event from their past as if it was currently happening. The mental tasks phase included assessing cognitive load through tasks such as the modified Stroop Colour-Word Task (SCWT) (Stroop, 1935), requiring participants to read colour names (red, green, and blue) printed in incongruous ink (e.g., the word RED appearing in blue ink). The difficulty was increased by asking participants first to read each word and then name the colour of the word. A sec-

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

#	Experimental task	Affective State
Social Exposure		
1	1.1 Neutral (Reference)	N
2	1.2 Baseline Description	N
3	1.3 Interview	S
Emotional Recall		
4	2.1 Neutral (Reference)	N
5	2.2 Recall stressful event	S
Mental Workload		
6	3.1 Reading words (Reference)	N
7	3.2 Stroop Colour-Word Test	S
8	3.3 PASAT task	S
Stressful Stimuli		
9	4.1 Relaxing video	R
10	4.2 Adventure video	S
11	4.3 Psych. pressure video	S

ond mental task used was the Paced Auditory Serial Addition Test (PASAT) (Gronwall, 1977), which is a neuropsychological test involving arithmetic operations to assess attentional processing. The stressful video phase included the presentation of 2-minute videos designed to induce low-intensity positive emotions (calming video) and stress (action scene from an adventure film, a scene involving heights to participants with moderate levels of acrophobia, a burglary/home invasion while the inhabitant is inside, car accidents etc.). Each participant gave their free and informed permission and the Research Ethics Committee of FORTH provided its approval for this study (approval no. 155/12-09-2022).

4 METHODS

4.1 Contrastive Learning Feature Extraction

MIL models usually use standard feature networks, such as C3D (Tran et al., 2015) or I3D (Carreira and Zisserman, 2017), trained on action detection datasets such as Kinetics-400. Such networks may not be well suited for the detection of medical abnormalities in facial video data. Given this limitation, we consider using a contrastive learning network that was trained on facial video data instead. In (Brügge et al., 2023), it was demonstrated that using this network, it was possible to extract distinguishing features for the medical task, despite being trained solely on data from healthy individuals. Applying the network requires the detection and tracking of facial landmarks. Thus,

for contrastive learning, it is necessary to use tailored transformations, such as flipping the landmark coordinates horizontally and global and local scaling, implemented by a multiplication of x - and y -coordinates by random factors.

4.2 Multiple Instance Learning

4.2.1 Motivation

Multiple Instance Learning is a learning approach that trains a model using weak labels at the video level to infer unknown labels at second or snippet level. The video data is divided into positive and negative bags. In our stress detection task, positive bags represent videos that contain at least one shorter video snippet showing stress behaviour and negative bags represent videos showing solely neutral or relaxed behaviour. Top- k MIL (Angles et al., 2021), (Li and Vasconcelos, 2015b), (Tian et al., 2021) identifies the top- k instances within each bag that are likely to be positive examples and uses this information to classify each instance in the bag. Due to the absence of second-wise labelled data for evaluation, we focus on improving classification performance at the bag level. At the same time, by using MIL we obtain a temporal instance segmentation, which improves explainability by providing insight into which snippets contain stress behaviour.

In stress detection, typically only a small proportion of video snippets exhibit stress behaviour, making the majority of the content appear normal. To effectively use these data for training, we consider not only the top- k snippets but also the bottom- k snippets within positive bags. We assume that the majority of snippets in positive bags show no signs of stress, allowing us to label bottom- k snippets as neutral instances. Incorporating these bottom- k snippets into the training dataset as normal instances could help to make better use of limited datasets and improve the robustness of the classifier.

4.2.2 Bottom-k Multiple Instance Learning

In our stress detection task, we cut the videos into non-overlapping sub-videos to form bags. Each bag contains a fixed number of features extracted using the contrastive learning feature network. With the pre-extracted features $\mathbf{F}_i \in \mathbb{R}^{T \times D}$ and the corresponding weak video-level binary stress label y_i , we denote the training dataset of weakly-labelled recordings as $\mathcal{D} = (\mathbf{F}_i, y_i)_{i=1}^{|\mathcal{D}|}$. D and T denote the feature size and the number of features in a single training video, respectively. The label y_i takes the value 0 if it shows

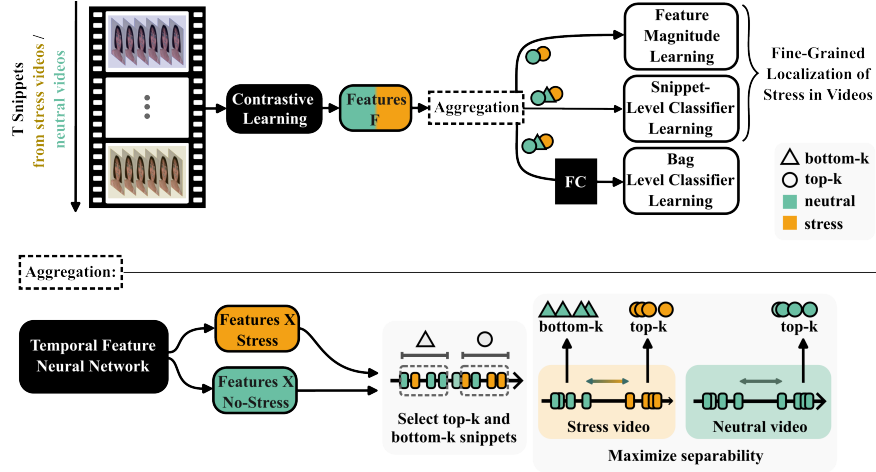


Figure 1: Overview over the proposed multiple instance learning approach for stress detection. Videos are divided into T snippets from which contrastive learning features are extracted. These features are input to another temporal feature neural network. Using the feature magnitude learning scheme, the separability of neutral and stress snippets is maximised. The resulting top- k and bottom- k snippets of stress videos and top- k snippets of neutral videos then serve as input to a snippet-level and a bag-level classifier. Bottom- k snippets of stress videos are labelled as neutral.

the subject during an experimental task that was assigned the affective state ‘‘N’’ or ‘‘R’’ and it takes the value 1 for the affective state ‘‘S’’.

An overview over our bottom- k MIL approach is given in Figure 1. We use a multi-head-attention temporal feature network $s_\theta : \mathcal{F} \rightarrow \mathcal{X}$ (see subsection 4.3, Figure 3 and Figure 2 for details) for the extraction of temporal features $\mathbf{X} = s_\theta(\mathbf{F})$ from the features \mathbf{F} . Based on these temporal features, a snippet-level classification network $f_\phi : \mathcal{X} \rightarrow [0, 1]^T$ is generating the binary classification whether a video snippet contains stress behaviour by $f_\phi(s_\theta(\mathbf{F}))$. Features of positive and negative snippets are denoted as $\mathbf{x}^+ \sim P_x^+(\mathbf{x})$ and $\mathbf{x}^- \sim P_x^-(\mathbf{x})$, as in (Tian et al., 2021). With $t = 1, \dots, T$, a snippet feature \mathbf{x}_t represents the t -th row in \mathbf{X} . A positive video \mathbf{X}^+ showing stress behaviour can contain snippets drawn from both $P_x^+(\mathbf{x})$ and $P_x^-(\mathbf{x})$ but negative videos \mathbf{X}^- showing normal behaviour can only contain snippets from $P_x^-(\mathbf{x})$. We also make the assumption that $\mathbb{E}[\|\mathbf{x}^+\|_2] \geq \mathbb{E}[\|\mathbf{x}^-\|_2]$ indicating that stress snippet features have larger magnitudes than non-stress snippet features.

The snippet-level classifier f_ϕ , the temporal feature network s_θ and the bag-level classifier c_ψ are trained jointly. The joint loss is given by

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

with N being the number of input sub-videos of length T extracted from one recording. The loss function

ℓ_{overall} combines a cross-entropy snippet classification loss ℓ_f , a feature separability loss function ℓ_s and a bag loss function ℓ_b . We outline the different loss terms below.

We use the feature separability loss ℓ_s from (Tian et al., 2021) to ensure that the feature magnitude correlates with the probability of a snippet feature being positive. The mean feature norm is calculated by

$$g_{\theta,k}(\mathbf{X}) = \max_{\Omega_k(\mathbf{X}) \subseteq \{\mathbf{x}_t\}_{t=1}^T} \frac{1}{k} \sum_{\mathbf{x}_t \in \Omega_k(\mathbf{X})} \|\mathbf{x}_t\|_2 \quad (2)$$

where $\Omega_k(\mathbf{X})$ is a subset of k snippets in $\{\mathbf{x}_t\}_{t=1}^T$. The separability loss ℓ_s is given by

$$\begin{aligned} \ell_s(s_\theta(\mathbf{F}_i), s_\theta(\mathbf{F}_j), y_i, y_j) = \\ \begin{cases} (|m - g_{\theta,k}(\mathbf{X}^+)| + g_{\theta,k}(\mathbf{X}^-))^2 & \text{if } y_i = 1, y_j = 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \quad (3)$$

where m is a pre-selected margin.

The classification cross-entropy loss ℓ_f is given by

$$\begin{aligned} \ell_f(f_\phi(s_\theta(\mathbf{F})), y) = \\ \sum_{\substack{\mathbf{x} \in \Omega_{k,\max}(\mathbf{X}) \\ \mathbf{x} \in \Omega_{k_b,\min}(\mathbf{X})}} -(y \log(f_\phi(\mathbf{x})) + (1-y) \log(1 - f_\phi(\mathbf{x}))). \end{aligned} \quad (4)$$

This loss is not only getting the top- k features from \mathbf{X}_i with the largest L2 norm as input but also the bottom- k features with the smallest L2 norm of all snippets in a positive bag. Top- k snippets are represented by the set $\Omega_{k,\max}(\mathbf{X})$ and bottom- k snippets are represented by the set $\Omega_{k_b,\min}(\mathbf{X})$. For the bottom- k features, the label y takes the value 0 because we assume that positive bags also contain no-stress snippets.

For calculating the bag-level loss, the features \mathbf{X} are input to a simple bag-classification head $c_\psi(\cdot)$ consisting of two fully-connected network layers and ReLU activation. The cross-entropy classification loss is given as

$$\begin{aligned} \ell_b(c_\psi(f_\phi(s_\theta(\mathbf{F}))), y) = \\ - (y \log(c_\psi(f_\phi(\tilde{\mathbf{X}}))) + (1 - y) \log(1 - c_\psi(f_\phi(\tilde{\mathbf{X}})))) \end{aligned} \quad (5)$$

where $\tilde{\mathbf{X}}$ contains all top- k and bottom- k features of \mathbf{X} .

4.3 Temporal Feature Network

For temporal feature extraction, we use a neural network employing multi-head attention and convolutions at different temporal scales. We therefore call this network Multi-Scale Multi-Head Attention Network (MSMHN). An overview of this network is given in Figure 3. The stress-related information is extracted at different temporal scales from the input features \mathbf{F} . As also done in (Tian et al., 2021), this is achieved by using dilated convolutions in the temporal direction of \mathbf{F} . The dilation factors of the three 1D-convolutional network branches are 1, 2 and 3 to capture both subtle short- and long-term facial expression patterns.

Each dilated convolution branch of the network is equipped with its own multi-head self-attention mechanism (Vaswani et al., 2017). Figure 2 gives an overview over the self-attention module. Multi-Head attention divides the attention mechanism into several parallel and individual heads to compute attention scores. Each head uses dot product attention, a process that calculates attention weights by computing the dot product between a query and key input vector.

After computing the dot product attention, the outputs from all heads are concatenated, added to the input sequence and then normalised to generate the final multi-head attention output. This self-attention mechanism enables the network to identify relevant data patterns and relationships at different temporal scales.

For feature fusion, the extracted multi-scale temporal features are concatenated and processed through a convolutional layer. Additionally, we employ a skip connection.

5 EXPERIMENTAL DETAILS

We trained the self-supervised feature network according to (Brügge et al., 2023) and applied it to our stress dataset. To train the MIL framework, we use

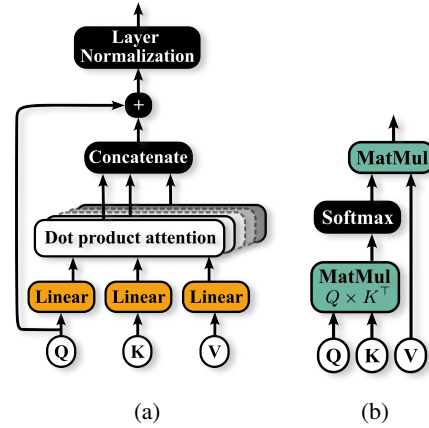


Figure 2: Network sub-modules of the Multi-Scale Multi-Head Network. The multi-head attention mechanism is shown in a), dot product attention in b). Query (Q), Key (K) and Value (V) are given by the output of the dilated convolutions (Figure 3).

data from one neutral/relaxed and one stressful video from the same experimental task, yielding a balanced dataset. This process is repeated for all task combinations, resulting in a separate network trained for each task combination. One bag was represented by $T = 30$ consecutive feature snippets of one subject performing a single task. Segments that form a bag were chosen without overlap.

We used Adam optimisation with a learning rate of 10^{-4} and a batch size of 32. We set the parameters $k = 10$ and $k_b = k$ for the bottom- k snippets. As bag classifier, we used a simple two-layer fully-connected (FC) neural network with 512 nodes in the hidden layer and ReLU activation. We train the MIL framework for 10 epochs.

To validate our approach, we used 10-fold cross-validation, where in each fold we excluded 5 subjects from the training set and used their data for the evaluation. In this way, we investigate the extent to which our approach generalises to unseen subjects.

6 RESULTS AND DISCUSSION

In this section, we present our results, which demonstrate the effectiveness of our proposed MIL method for stress detection. In the following, we summarise our experiments and report on the results for different networks and training schemes. We listed the results of all experiments in subsection 2.

3D ResNet-18 Trained with Dense Labels. In a first experiment, we trained a 3D ResNet-18 (Hara et al., 2017) as baseline model for fully supervised snippet-level training. As in all following experiments, train-

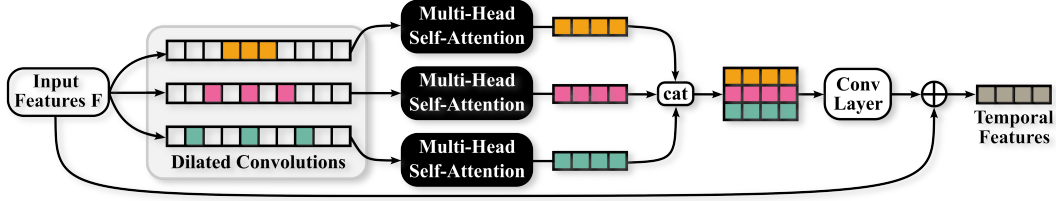


Figure 3: Network architecture of the Multi-Scale Multi-Head Network for temporal feature extraction. Multi-Scale temporal features are extracted using dilated convolutions. These are then input to a multi-head self-attention mechanism, shown in subsection 2 and concatenated afterwards. The concatenated self-attention features are input to another convolutional layer for feature fusion. The original features are added through a skip connection.

Table 2: Stress classification bag-level accuracy (ACC) and F1 Score (F1) for different network architectures and learning schemes, averaged over all experimental tasks. 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 4.2.1.

Model	MIL	ACC (%)	F1 (%)
3D ResNet-18 + Dense Label	✗	77.28 ± 16.92	76.34 ± 23.45
3D ResNet-18	✗	83.38 ± 12.88	81.32 ± 21.43
MTN (Tian et al., 2021)	✗	86.14 ± 12.89	82.68 ± 12.43
MTN (Tian et al., 2021)	✓	93.19 ± 5.21	93.57 ± 4.71
MTN + Bottom- k	✓	94.17 ± 5.19	94.28 ± 5.13
MSMHN	✓	95.09 ± 4.77	95.22 ± 4.63
MSMHN + Bottom- k	✓	95.46 ± 4.37	95.49 ± 4.77

Table 3: Stress classification bag-level accuracy (ACC) and F1 score (F1) for the 7 different stress task combinations. As network we used the best-performing MSMHN architecture and trained it using the proposed bottom-score MIL approach according to Table 2.

	Task	ACC (%)	F1 (%)
Social Exposure	1.2 vs. 1.3	97.78 ± 1.57	97.78 ± 1.55
Emotional Recall	2.1 vs. 2.2	96.77 ± 2.80	96.82 ± 2.63
Mental Workload	3.1 vs. 3.2	94.03 ± 3.46	94.34 ± 3.18
Mental Workload	3.1 vs. 3.3	95.19 ± 3.14	95.35 ± 3.00
Stressful Stimuli	4.1 vs. 4.2	97.35 ± 2.08	97.29 ± 2.09
Stressful Stimuli	4.1 vs. 4.3	91.68 ± 7.15	91.35 ± 7.64

ing was performed on the extracted contrastive learning features to ensure comparability and to analyse the effect of MIL on the results. For this baseline, we labelled each snippet with the bag-label to get a densely-labelled dataset and fed these snippets to the ResNet individually. This approach resulted in an accuracy of 77.28 % and an F1 score of 76.34 %.

3D ResNet-18 Trained with Bag-Level Labels. In a second experiment, we used the same 3D ResNet-18 architecture but divided the video into 30 s segments and assigned a single label to each segment. All 30 snippets of one bag were fed into the FC bag classifier at once. With this approach, accuracy and F1 score increased to 83.38 % and 81.32 %, respectively.

MTN. We followed the same training strategy using

the MTN architecture from (Tian et al., 2021), since it has proven useful in extracting temporal features from video data features. The model combines parallel dilated convolutions with dilation factors up to 4, an attention network branch and residual connections. Again, all extracted features were input to the bag classifier. Using this architecture improved performance, yielding an accuracy of 86.14 % and an F1 score of 82.68 %.

MTN Trained Using MIL. We used the same MTN architecture and trained the network using top- k MIL as described in 4.2, but without using the bottom scores. Using MIL again be improved the accuracy and F1 score to 93.19 % and 93.57 %, respectively.

MTN Trained Using MIL with Bottom Scores. Next, we incorporated bottom- k features from stress videos into the snippet- and bag-level classification losses ℓ_f and ℓ_b in training the MTN. Integrating bottom- k features has led to further improvements, resulting in an accuracy of 94.17% and an F1 score of 94.28 %.

MSMHN Trained Using MIL. For a direct comparison, we train the proposed MSMHN, described in subsection 4.3 using MIL without using bottom- k features. This improved the accuracy and F1 score by approximately one percent compared to the best results obtained with the MTN, leading to 95.09 % and 95.22 %, respectively.

MSMHN Trained Using MIL with Bottom Scores. Finally, we integrated bottom- k scores for bag-level training also in the MIL training of the MSMHN. The proposed model performed best. It achieved an accuracy of 95.46 % and an F1 score of 95.49 %, outperforming all previously considered methods. We performed a statistical analysis on the results. Shapiro-Wilk tests confirmed that the data were normally distributed. A paired t-test showed statistically significant differences in the performance measures compared to MSMHN using no bottom scores. Specifically, the F1 scores showed $t(59) = 2.324, P = 0.023$, significant at $P < 0.05$, and the accuracy showed

$t(59) = 3.751, P = 0.0004$, significant at $P < 0.001$.

A possible explanation is that models are more likely to overfit on small datasets and may learn small differences in the stress videos that are not indicative of stress. The addition of the bottom snippets from stress videos as neutral snippets contributes to the diversity of the dataset, as characteristics such as head posture may differ between neutral and stress videos.

We have listed the results of the best performing model for all task combinations in Table 3. The table shows that the classification yielded the best results for the subjects performing a baseline description vs. being involved interview and watching a relaxing video vs. an adventure video. In this setting, the classification accuracies are 97.78 % and 97.35 % and the F1 scores are 97.79 % and 97.29 %, respectively. The lowest accuracy and F1 score were obtained by classifying videos of subjects watching a relaxing video vs. a psychological pressure video with an accuracy of 91.68 % and an F1 score of 91.35 %.

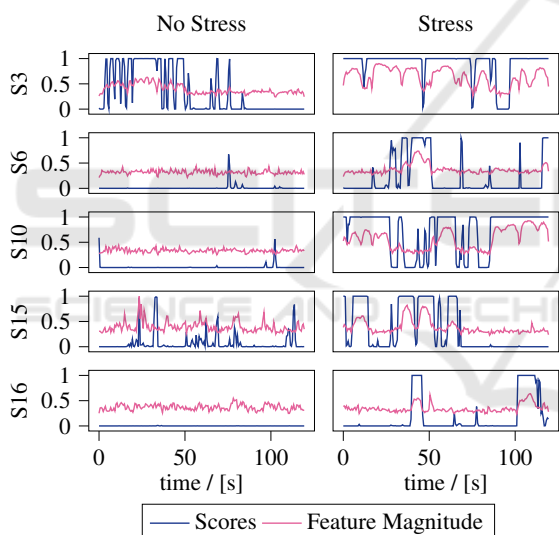


Figure 4: Snippet classifier scores and feature magnitudes during the neutral task 4.1 (left) and stressful stimuli task 4.3 (right) for five exemplary subjects and a time span of 2 min.

We provide exemplary sequences of facial feature magnitudes and network scores in relaxed (task 4.1) and stressed (task 4.3) videos for five subjects in Figure 4. The plots show the feature magnitudes and snippet classifier scores over time, highlighting the time steps the network focuses on. It therefore contributes to the explainability of our approach and could provide insights into stress dynamics in facial videos. The figure shows that the norm of the features increases with a high prediction score. This is more pronounced for the stress tasks than for the neutral tasks. Additionally, the model highlights short seg-

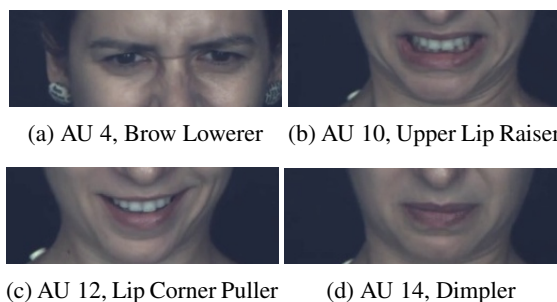


Figure 5: Action Units that had a significant correlation with the stress prediction scores of the MIL neural network.

ments in the neutral tasks as seen in the first row in Figure 4.

Correlations Between Predictions and Action Unit Time Series.

To further analyse what the network focuses on, we use 17 facial action unit (AU) time series for each subject and task and calculate their correlation with the network predictions and feature magnitudes. The correlations are false detection rate corrected and shown as box plots in Figure 6 the appendix. The action units that correlate with the prediction vary depending on the task. AU 14 (Dimpler, see Figure 5d) is strongly pronounced in most tasks. With the exception of task 4.2 vs. 4.3, the time series of predictions and features of more than 10 people show at least a moderate correlation of 40 % with this AU. In the neuropsychological test tasks (Mental Workload, tasks 3.1 vs. 3.2 and 3.1 vs. 3.3), we also found that AUs 10, and 12 (Upper Lip Raiser and Lip Corner Puller, see Figure 5b and Figure 5c) are highly pronounced. In addition, AU 4 (Brow Lowerer, Figure 5a) is particularly present in many test subjects in all tasks except 1.2 vs. 1.3 and 2.1 vs. 2.2.

Limitations. While the results demonstrate robust stress detection capabilities, it is important to mention potential limitations and open research questions.

First, without snippet-level labels, it is not possible to assess whether the network exclusively focuses on stress indicators. This is particularly evident in the mental load task, where subjects tend to smile after making errors. Although smiling is not inherently indicative of stress, the network may associate it with the subject making errors that occur during the stress task. This is supported by the greater prevalence of subjects with at least moderate correlations of AUs 6, 10 and 12 with the network predictions in the tasks involving neuropsychological tests compared to the other task combinations. These AUs are the cheek raiser, the upper lip raiser and the lip corner puller, which are typically activated when a person smiles.

Secondly, we would like to mention that determining the optimal choice of the parameter k is an

ongoing challenge, as its choice is highly dependent on the considered dataset. Consequently, the parameter used might not necessarily be ideal for different datasets and should be chosen based on the frequency of the anomalies to be detected. Also, this study did not focus on improving snippet feature extraction, and there may be other feature extraction models that further improve the classification. However, it has been shown in (Brügge et al., 2023) that the applied model performs well for facial video data.

Finally, it should be noted that all videos were recorded in a controlled recording environment, which ideally should also apply to the data on which the model is evaluated. However, the use of contrastive learning as a feature extractor mitigates this point, as contrastive learning introduces invariance to various influences. In addition, because contrastive learning is applied to landmarks, influences such as appearance, lighting and background play a minor role, as long as the landmark detection is robust.

7 CONCLUSION

The results showed that a high stress detection accuracy was achieved when MIL was applied to the facial video data of subjects performing different neutral and stressful tasks. In an ablation study, we successively motivated the components of our approach by evaluating the use of MIL, the temporal feature network and the integration of bottom scores. In our dataset, where we expect anomalous events to be scarce, stress detection using neural networks can benefit from a MIL training scheme where the instances most likely to be anomalous are considered for classification. The proposed MSMHN also led to improved results. Further improvement of this MIL baseline could be achieved by including segments that are unlikely to contain stress behaviour, even though they are sampled from a video taken during a stressful task. Using the combination of our proposed modifications to top- k MIL, the stress detection accuracy and F1 scores averaged over all experimental tasks were 95.46 % and 95.49 %, respectively.

The use of MIL simultaneously provides valuable insights into which snippets contribute most to the classification of stress behaviour. We used correlation analysis to identify the action units that are predominantly activated in these critical snippets. In future work, we aim to further increase this explainability by highlighting specific facial regions that play a key role in stress classification to better understand the manifestation of stress in facial expressions.expressions.

ACKNOWLEDGEMENTS

This work was supported by the BMBF (16KISA057).

REFERENCES

- Angles, B., Jin, Y., Kornblith, S., Tagliasacchi, A., and Yi, K. M. (2021). MIST: Multiple instance spatial transformer. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2412–2422. IEEE.
- Bevilacqua, F., Engström, H., and Backlund, P. (2018). 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.
- Bobade, P. and Vani, M. (2020). Stress detection with machine learning and deep learning using multimodal physiological data. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 51–57.
- Brügge, N. S., Mohammadi, E., Münchau, A., Bäumer, T., Frings, C., Beste, C., Roessner, V., and Handels, H. (2023). 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)*.
- Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, Los Alamitos, CA, USA. IEEE Computer Society.
- Chrousos, G. P. (2009). Stress and disorders of the stress system. *Nature reviews endocrinology*, 5(7):374.
- Daudelin-Peltier, C., Forget, H., Blais, C., Deschênes, A., and Fiset, D. (2017). The effect of acute social stress on the recognition of facial expression of emotions. *Scientific Reports*, 7(1):1036.
- Dundar, M. M., Badve, S., Raykar, V. C., Jain, R. K., Sertel, O., and Gurcan, M. N. (2010). A multiple instance learning approach toward optimal classification of pathology slides. In *2010 20th International Conference on Pattern Recognition*, pages 2732–2735. ISSN: 1051-4651.
- Feng, J.-C., Hong, F.-T., and Zheng, W.-S. (2021). MIST: Multiple instance self-training framework for video anomaly detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14004–14013. IEEE.
- Garg, P., Santhosh, J., Dengel, A., and Ishimaru, S. (2021). Stress detection by machine learning and wearable sensors. In *26th International Conference on Intelligent User Interfaces - Companion, IUI '21 Companion*, pages 43–45. Association for Computing Machinery.

- Gavrilescu, M. and Vizireanu, N. (2019). Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors*, 19(17):3693. Publisher: Multidisciplinary Digital Publishing Institute.
- Geng, S., Jia, S., Qiao, Y., Yang, J., and Jia, Z. (2015). Combining CNN and MIL to assist hotspot segmentation in bone scintigraphy. In Arik, S., Huang, T., Lai, W. K., and Liu, Q., editors, *Neural Information Processing*, Lecture Notes in Computer Science, pages 445–452. Springer International Publishing.
- Giannakakis, G., Grigoriadis, D., Giannakaki, K., Simanti-raki, O., Roniotis, A., and Tsiknakis, M. (2019). Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*.
- Giannakakis, G., Koujan, M. R., Roussos, A., and Marias, K. (2020). Automatic stress detection evaluating models of facial action units. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 728–733. IEEE.
- Giannakakis, G., Pedititis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., Marias, K., and Tsiknakis, M. (2017). Stress and anxiety detection using facial cues from videos. *Biomed. Signal Process. Control.*, 31:89–101.
- Giannakakis, G., Roussos, A., Andreou, C., Borgwardt, S., and Korda, A. I. (2025). Stress recognition identifying relevant facial action units through explainable artificial intelligence and machine learning. *Computer Methods and Programs in Biomedicine*, 259:108507.
- Gronwall, D. (1977). Paced auditory serial-addition task: a measure of recovery from concussion. *Perceptual and motor skills*, 44(2):367–373.
- Hara, K., Kataoka, H., and Satoh, Y. (2017). Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3154–3160.
- Hasani, B. and Mahoor, M. H. (2017). Facial expression recognition using enhanced deep 3d convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 30–40.
- Hosseini, S., Katragadda, S., Bhupatiraju, R. T., Ashkar, Z., Borst, C., Cochran, K., and Gottumukkala, R. (2021). A multi-modal sensor dataset for continuous stress detection of nurses in a hospital.
- Jeon, T., Bae, H. B., Lee, Y., Jang, S., and Lee, S. (2021). Deep-learning-based stress recognition with spatial-temporal facial information. *Sensors*, 21(22):7498.
- Kandemir, M. and Hamprecht, F. A. (2015). Computer-aided diagnosis from weak supervision: a benchmarking study. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, 42:44–50.
- Kandemir, M., Zhang, C., and Hamprecht, F. A. (2014). Empowering multiple instance histopathology cancer diagnosis by cell graphs. In Golland, P., Hata, N., Barillot, C., Hornegger, J., and Howe, R., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, Lecture Notes in Computer Science, pages 228–235. Springer International Publishing.
- Korda, A. I., Giannakakis, G., Ventouras, E., Asvestas, P. A., Smyrnis, N., Marias, K., and Matsopoulos, G. K. (2021). Recognition of blinks activity patterns during stress conditions using cnn and markovian analysis. *Signals*, 2(1):55–71.
- Kumar, S., Iftexhar, A. S. M., Goebel, M., Bullock, T., Maclean, M., Miller, M., Santander, T., Giesbrecht, B., Grafton, S., and Manjunath, B. (2021). Stressnet: Detecting stress in thermal videos. pages 998–1008.
- Li, H., Yang, F., Xing, X., Zhao, Y., Zhang, J., Liu, Y., Han, M., Huang, J., Wang, L., and Yao, J. (2021). 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*, pages 529–539. Springer.
- Li, R. and Liu, Z. (2020). Stress detection using deep neural networks. *Z.*
- Li, S., Liu, F., and Jiao, L. (2022). Self-training multi-sequence learning with transformer for weakly supervised video anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1395–1403.
- Li, W. and Vasconcelos, N. (2015a). Multiple instance learning for soft bags via top instances. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4277–4285.
- Li, W. and Vasconcelos, N. (2015b). Multiple instance learning for soft bags via top instances. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4277–4285. IEEE.
- Luo, W., Liu, W., and Gao, S. (2017). A revisit of sparse coding based anomaly detection in stacked RNN framework. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 341–349. IEEE.
- Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. ISSN: 1063-6919.
- Naegelin, M., Weibel, R. P., Kerr, J. I., Schinazi, V. R., La Marca, R., von Wangenheim, F., Hoelscher, C., and Ferrario, A. (2023). An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *Journal of Biomedical Informatics*, 139:104299.
- Siam, A. I., Gamel, S. A., and Talaat, F. M. (2023). Automatic stress detection in car drivers based on non-invasive physiological signals using machine learning techniques. 35(17):12891–12904.
- Sikka, K., Dhall, A., and Bartlett, M. (2013). Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.
- Sikka, K., Dhall, A., and Bartlett, M. S. (2014). Classification and weakly supervised pain localization using

- multiple segment representation. *Image and Vision Computing*, 32(10):659–670.
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18(6):643.
- Sultani, W., Chen, C., and Shah, M. (2018). Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488. IEEE.
- Tian, Y., Pang, G., Chen, Y., Singh, R., Verjans, J. W., and Carneiro, G. (2021). Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4955–4966. IEEE.
- Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., and Carneiro, G. (2022). 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*, pages 88–98. Springer-Verlag.
- Tong, T., Wolz, R., Gao, Q., Hajnal, J., and Rueckert, D. (2013). 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*, 16:599–606.
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497. IEEE.
- Turcan, E., Muresan, S., and McKeown, K. (2021). 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*, pages 2895–2909. Association for Computational Linguistics.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Viegas, C., Lau, S.-H., Maxion, R., and Hauptmann, A. (2018). Towards independent stress detection: A dependent model using facial action units. *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6.
- Wan, B., Fang, Y., Xia, X., and Mei, J. (2020). Weakly supervised video anomaly detection via center-guided discriminative learning. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.
- Wang, Y., Ma, J., Hao, B., Hu, P., Wang, X., Mei, J., and Li, S. (2020). Automatic depression detection via facial expressions using multiple instance learning. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1933–1936.
- Wu, P., Liu, J., Shi, Y., Sun, Y., Shao, F., Wu, Z., and Yang, Z. (2020). 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*, pages 322–339. Springer-Verlag.
- Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., and Chang, E. I.-C. (2014a). 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)*, pages 1626–1630. ISSN: 2379-190X.
- Xu, Y., Zhu, J.-Y., Chang, E., Lai, M., and Tu, Z. (2014b). Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18:591–604.
- Xu, Y., Zhu, J.-Y., Chang, E., and Tu, Z. (2012). Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 964–971. ISSN: 1063-6919.
- Zhang, H., Feng, L., Li, N., Jin, Z., and Cao, L. (2020). Video-based stress detection through deep learning. *Sensors*, 20(19):5552.
- Zhang, J., Qing, L., and Miao, J. (2019). Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4030–4034.
- Zhang, J., Yin, H., Zhang, J., Yang, G., Qin, J., and He, L. (2022). Real-time mental stress detection using multi-modality expressions with a deep learning framework. *Frontiers in Neuroscience*, 16:947168.
- Zhong, J.-X., Li, N., Kong, W., Liu, S., Li, T. H., and Li, G. (2019). 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)*, pages 1237–1246. IEEE.

APPENDIX

Correlations

In Figure 6 we show the correlations between network predictions and action unit intensities for the different task combinations. All correlations have been corrected for false detection rates. The objective was to identify the action units that were most pronounced in the time steps that correspond to the highest network predictions to introduce explainability into our method.

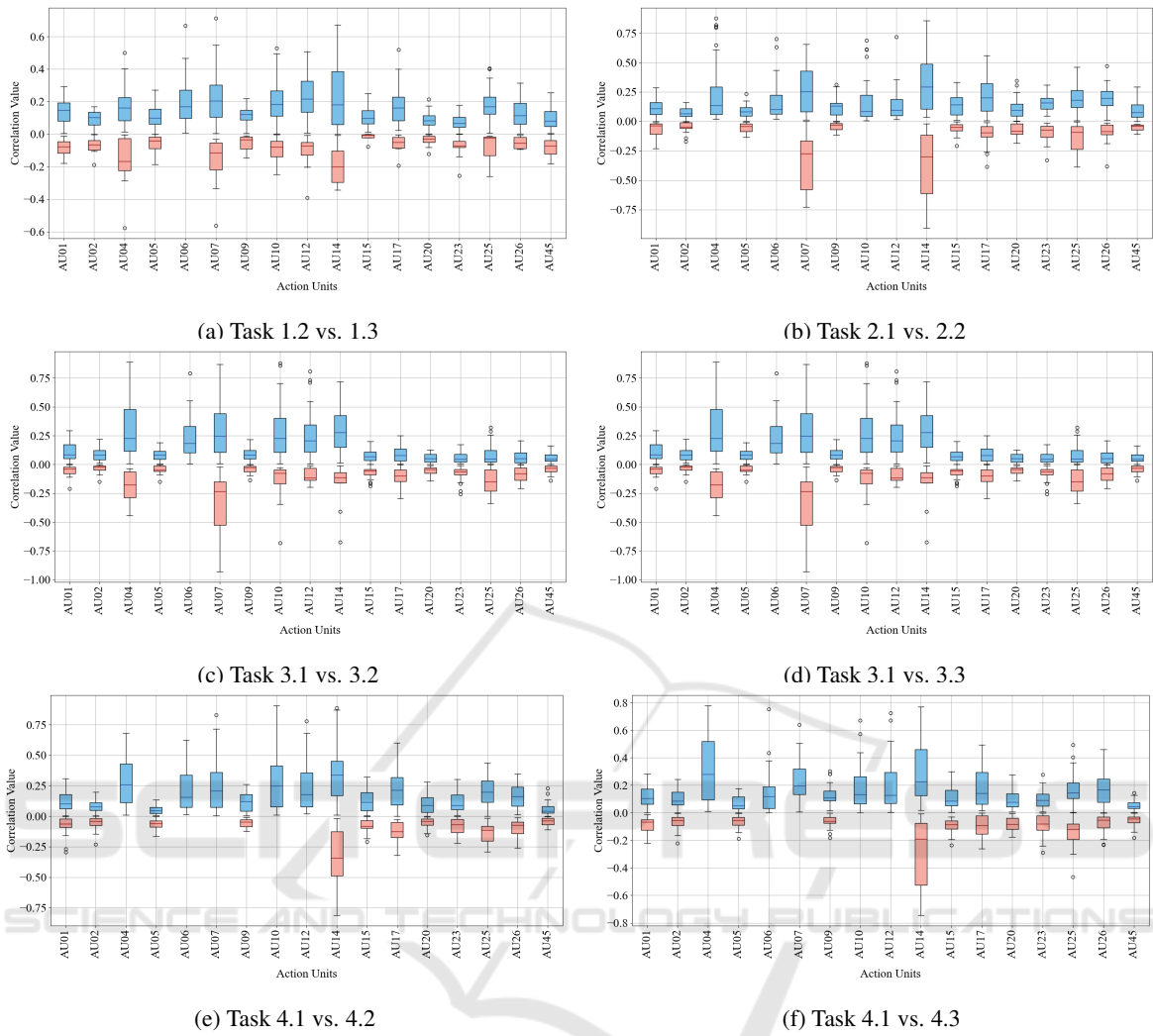


Figure 6: Box plots showing the correlations between the network predictions and the action unit intensities for the different task combinations. Positive (blue) and negative (red) correlations are shown one above the other in two different box plots.