Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data

Debasmita Ghose*
Yale University
debasmita.ghose@yale.edu

Oz Gitelson*
Yale University
oz.gitelson@yale.edu

Brian Scassellati
Yale University
brian.scassellati@yale.edu

Figure 1: Key frames from our sample video clips from our MultiAffectStress (MAS) dataset with a) video labeled as "Stress" containing instances of Facial Stress, Fidgeting, Vocal Stress and Sentiment Stress and, b) video labeled as "No Stress"

ABSTRACT

Stress detection in real-world settings presents significant challenges due to the complexity of human emotional expression influenced by biological, psychological, and social factors. While traditional methods like EEG, ECG, and EDA sensors provide direct measures of physiological responses, they are unsuitable for everyday environments due to their intrusive nature. Therefore, using non-contact, commonly available sensors like cameras and microphones to detect stress would be helpful. In this work, we use stress indicators from four key affective modalities extracted from audio-visual data: facial expressions, vocal prosody, textual sentiment, and physical fidgeting. To achieve this, we first labeled 353 video clips featuring individuals in monologue scenarios discussing personal experiences, indicating whether or not the individual is stressed based on our four modalities. Then, to effectively integrate signals from the four modalities, we extract stress signals from our audio-visual data using unimodal classifiers. Finally, to explore how the different modalities would interact to predict if a person is stressed, we compare the performance of three multimodal fusion methods: intermediate fusion, voting-based late fusion, and learning-based late fusion. Results indicate that combining multiple modes of information can effectively leverage the strengths of different modalities and achieve an F1 score of 0.85 for binary stress detection. Moreover, an ablation study shows that the more modalities are integrated, the higher the F1 score for detecting stress across all fusion techniques, demonstrating that our selected modalities possess complementary stress indicators.

CCS CONCEPTS

• **Human-centered computing** → **Ambient intelligence**.

KEYWORDS

stress detection; affective computing; multimodal fusion

ACM Reference Format:
Debasmita Ghose, Oz Gitelson, and Brian Scassellati. 2024. Integrating Multimodal Affective Signals for Stress Detection from Audio-Visual Data. In ,. ACM, New York, NY, USA, 11 pages. https://doi.org/https://doi.org/10.1145/3678957.3685717

*Both authors contributed equally to the paper

ICMI '24, November 4–8, 2024, San Jose, Costa Rica,

ACM ISBN ISBN HERE
https://doi.org/https://doi.org/10.1145/3678957.3685717

1 INTRODUCTION

Humans express stress in various ways, influenced by biological, psychological, and social factors. Stress significantly impacts human decision-making and affects how people experience the world around them [59]. In this regard, it is crucial to develop systems to detect stress. However, this is a challenging problem to solve, as even humans can struggle to accurately determine whether someone is stressed [56].

Researchers have used different strategies to identify stress. Typically, physiological data extracted from EEG, EDA, and Near Infrared Spectroscopy sensors are used to detect stress and have shown promising results in predicting emotional state or mental health conditions [46]. However, these systems rely on complex sensors that require physical contact with the person at all times to monitor their stress levels, making them impractical to deploy in uncontrolled environments.

An alternative stress detection method would be using non-contact sensors, such as cameras and microphones. Audio-visual data from these sensors could potentially be used to extract stress-related indicators. Some more apparent coarse-level indicators could be extracted from analyzing people's facial expressions, physical gestures, vocal prosody, and speech sentiment [18, 35, 54]. Some other subtle fine-grained stress-related indicators could be obtained from reasoning about people's shifts in eye gaze [32], speech rate variations [35], speech pauses [29], vocal tremors [79], and breathing patterns [52]. However, much of the research on stress detection from audio-visual sensors has only considered systems that use a single coarse-level indicator like facial expressions [6, 23, 90] voice [13], sentiment [28, 57, 65], and gestures [25, 47] to detect stress. More recently, efforts have been made to explore the interplay of vocal prosody with gesture [43] and facial expressions [2] for stress detection. However, these methods often overlook the complementary stress-related cues that can be obtained from jointly reasoning over more stress indicators. To address this gap, this paper proposes a multimodal fusion approach to stress detection that combines four coarse-level modalities: facial expressions, vocal prosody, physical gestures, and a person's speech sentiment. We chose to exclude the more fine-grained indicators in this work as determining if a person is stressed by analyzing those subtle features would require careful individual-level calibration between the times when they are stressed and when they are not, as these indicators might considerably vary between individuals.

Detecting stress accurately is inherently challenging due to multiple factors that influence an individual's expression of stress, compounded by the fact that stress is inherently multimodal. Simpler approaches that focus on a single modality often do not capture the full complexity of a person's stress responses. Moreover, the development of multimodal stress recognition is hampered by the lack of datasets specifically labeled for stress detection. Most stress recognition research has focused on emotion recognition rather than stress detection. This is because the datasets used to train stress detection models, such as SEWA [40], Aff-Wild2 [39], OMG-Emotion [8], and MUSE [30], are emotion recognition datasets. Hence, stress detection models have been trained using emotion recognition labels as a proxy for stress levels, which makes stress detection less accurate [89]. Moreover, these datasets do not contain labels for the four modalities our work focuses on. Therefore, we sample short video clips from YouTube and label them as *stress* or *no stress*. For clips with indicators of stress we label sentiment stress, vocal prosodic intonation stress, facial stress, and fidgeting. We call this dataset the MultiAffectStress (MAS) dataset.

To develop a multimodal stress detection model trained on our MultiAffectStress dataset, we used four individual classifiers to extract signals from each modality that we had labeled. We then combined the outputs of these classifiers using three different multimodal fusion techniques: intermediate fusion, voting-based late fusion, and learning-based late fusion. Our models were trained, validated, and evaluated on the MAS dataset. The best results were achieved by using all four modalities, and we obtained a maximum F1 score of 0.85. Finally, we conducted an ablation study to determine the contribution of each modality to our models' performance. Results showed that as each modality was added, the F1 scores of stress detection models increased for all three fusion techniques. This demonstrates that the modalities are complementary when performing stress detection from audio-visual data.

2 RELATED WORK

2.1 Stress Detection

Traditional stress detection methods have primarily relied on contact sensors, such as electroencephalograms (EEG) [34, 46, 81], functional near-infrared spectroscopy (fNIRS) [4, 33, 77], electrocardiograms (ECG) [36, 38, 66], and electrodermal activity (EDA) [14, 34, 74] sensors. However, these methods pose practical challenges for stress monitoring in everyday, non-clinical environments due to their intrusive nature and the requirement for physical contact with the individual [12, 58].

An emerging body of research has explored the potential of non-contact sensors, including thermal cameras [11, 41, 61, 88], ultrawideband radars [45, 76], and mobile phone sensors like accelerometers and gyroscopes [49, 75, 84] to detect physiological changes associated with stress without the need for direct contact with the human body. However, the deployment of such sensors in everyday situations remains limited by their accessibility, cost, and the complexity of their operation in uncontrolled environments. Therefore, using non-invasive audio and visual data to assess stress indicators is becoming increasingly popular [27].

Prior studies in the area of stress detection from audio-visual data have often focused on singular modalities, such as sentiment analysis [28, 57, 65], facial expression recognition [6, 23, 90], prosodic intonation in people's voice [13] or gesture analysis [16, 47, 50, 73, 86]. More recently, efforts have been made to explore the interplay of vocal prosody with gesture [43] and facial expressions [2] for stress detection. However, such studies are frequently focused in the context of acted or simulated stress [7, 13, 15, 20, 42, 48, 83], which may not accurately reflect genuine stress responses to real-world situations [31].

Inspired by the success of multimodal fusion techniques for extracting insights from multiple complementary data sources [24, 60, 87], we propose to leverage the multimodal nature of stress indicators to perform stress detection. We use monologue-style real-world human conversational videos and integrate signals extracted

from people's facial expressions, vocal intonation, sentiment analysis, and physical gestures like fidgeting to demonstrate the use of different multimodal fusion strategies [22] like intermediate fusion and late fusion. We show that our multimodal fusion model with all four stress signals outperforms models trained with fewer modalities, demonstrating that these modalities complement each other in building a robust stress detector that can be used for continuous, non-invasive stress monitoring outside clinical settings.

2.2 Audio-Visual Stress Detection Datasets

Recent years have seen the publication of several audio-visual emotion recognition datasets. However, these existing emotion recognition datasets often lack direct stress-related labels. Many well-established datasets such as SEWA [40], Aff-Wild2 [39], OMG-Emotion [8], and MUSE [30] contain either arousal/valence measures or categorical emotions like happiness, sadness, anger, disgust, fear and surprise. While these labels may be useful for emotional recognition, they only act as a derived proxy for stress [89], which may limit the accuracy of stress detection systems trained on them.

Many datasets used for training stress detection systems are not multimodal or connect only two modalities. Some datasets provide a sequence of images from videos without providing any audio data, leaving out important stress-related indicators that can be potentially captured through audio, and instead rely on physiological data or performance on a game [17, 44, 92]. Other datasets may include audio but do not provide enough video information [9, 19, 37]. A dataset by Ringeval et al. [64] includes video with audio, but only from the shoulders up, which prevents analysis of important signals related to pose and fidgeting.

Another common problem with a lot of related datasets is poor Inter-Rater Reliability (IRR) scores [78, 80, 83], as it is sometimes difficult even for humans to assess the ground-truth emotions shown by a person in a video [56]. Moreover, some stress-detection datasets have an imbalanced number of data points between stressed and non-stressed categories [10, 70]. These issues make it challenging to train reliable models on such datasets.

Our work proposes the MultiAffectStress (MAS) dataset that contains monologue-style real-world human conversational videos curated from two YouTube channels where people discuss stressful life experiences. Our dataset contains audio-visual data, and for each video clips, our labels contain direct labels for stress with a high Inter-Rater Reliability score (Cohen's Kappa = 0.85). For each video, we label no stress or one or more stress indicators - facial stress, vocal stress, sentiment stress, or fidgeting. Closest to our work, Lefter et al. [43] have published a dataset that contains videos with audio and has stress labeled on a five-point scale based on just gestures and voice. However, the videos in their dataset are composed of clips of actors enacting stressful scenarios, which may not accurately convey real-world emotions [31].

3 MULTI-AFFECT-STRESS (MAS) DATASET

In this work, we curated a collection of 353 video clips. These clips were sourced from two YouTube channels - "Keep it 100" playlists created by the channel *The Cut* [1] and videos from *Soft White Underbelly (SWU)* [2]. We selected these channels because they feature monologues of people recalling their stressful life experiences against a neutral background in a well-lit room while sitting on an interview stool, as seen in Fig. 1. We chose to use processed videos with proper lighting and simple backgrounds to minimize environmental factors that could interfere with stress detection models. Also, the use of stable cameras for our dataset helps maintain a consistent framing and focus on the subjects. Each video on the SWU channel features an interview with one person for the entire duration, while each video on the Cut channel features interviews with 100 people, one person at a time. We extracted clips from 12 videos (containing 112 clips) from The Cut and 121 videos (containing 241 clips) from SWU. In each video, only the upper body of a person is visible. Then, we used a two-step data annotation process:

(1) *Clip Selection:* Two researcherS with a background in Computer Science selected videos from two YouTube channels in reverse chronological order of publication. Then, for each video selected from the two channels, short video clips were selected using the Label Studio [82] video annotation tool. For the SWU channel, a video was deemed useful if it showed at least one instance with no stress signs and another that displayed one or more of the following stress indicators: facial stress, sentiment stress, vocal stress, or fidgeting. Facial stress was recognized when a person appeared angry, scared or was seen crying. If a person talked about a sad or frustrating incident or seemed angry, the clip was labeled as containing sentiment stress. Vocal stress was identified if a person's voice was shaking or heavy during a conversation. If a person displayed self-comforting gestures like touching their face repeatedly, the clip was considered to contain fidgeting. Since The Cut channel's videos consisted of interviews with 100 people in a single video, the researchers only included each person's video clip once in the entire dataset if it contained any of the above indicators of stress. We selected videos that showed clear differences in individuals' presentations or discussions under stressful and non-stressful conditions. Videos without discernible variations in expressions, tone, or behavior during stressful versus non-stressful situations were eliminated. The duration of the video clips ranged between 2.9 and 25.5 seconds, with an average of 9.8 ± 3.5 seconds. Since one of our modalities for stress detection is the contents of a person's speech, it was important to include their complete statements without truncating their sentences to limit our dataset to videos of fixed lengths.

(2) *Stress Indicator Annotation:* After the clips were selected, the same two researchers used the Label Studio [82] tool to annotate each video clip with one or more indicators of stress (sentiment stress, vocal stress, fidgeting or facial stress) or no stress. Each video clip was annotated with either *stress* or *no stress* labels as shown in Fig. 1. If a video was annotated with the label *stress*, it was additionally annotated with one of the four stress indicators. The Inter-Rater Reliability (IRR) score calculated using Cohen's Kappa method for labeling speech sentiment stress was 0.85, facial stress was 0.7, fidgeting was 0.56, and vocal prosodic stress was 0.42 between the two researchers. Overall, the IRR for stressed or non-stressed states was 0.85 indicating an excellent agreement.

[1] https://www.youtube.com/@cut

[2] https://www.youtube.com/@SoftWhiteUnderbelly

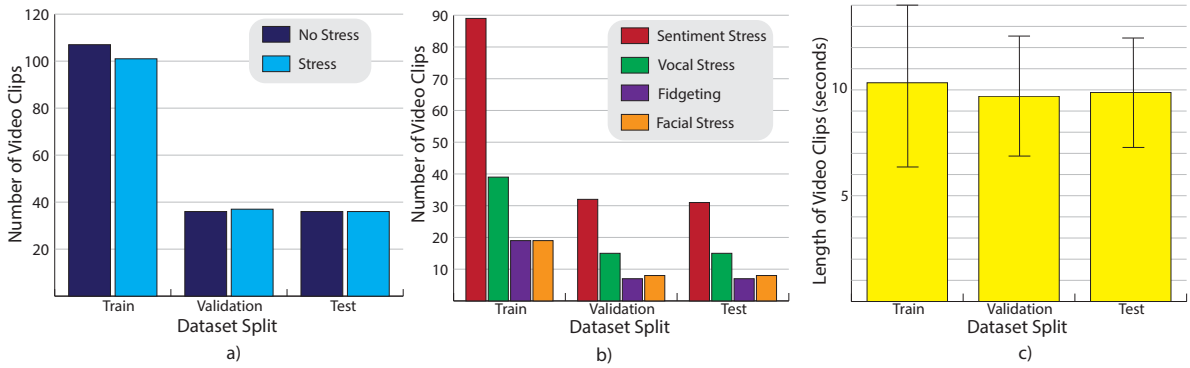

Figure 2: Dataset video clip statistics. a) Distribution of "Stress" and "No Stress" classes in the train, validation, and test splits of our MAS dataset. b) Distribution of each unimodal stress indicator in our MAS dataset's train, validation, and test split. Note that a given video can contain more than one stress indicator. c) Mean and standard deviation of the length of videos in the train, validation, and test split of our MAS dataset.

For benchmarking stress detection models, the dataset was split into three parts: 60% of the data for training, 20% for validation, and 20% for testing. The 353 clips in the dataset were distributed randomly across these splits based on their labels to maintain an even distribution of label variations and mean length of the video clips across each split. This distribution is shown in Figure 2. To prevent data leakage across the three splits, we ensured that only one stressed and one non-stressed clip from each video in the SWU channel and that no individual from a video in The Cut channel appeared in more than one clip across the splits. This dataset has been made publicly available [3] [4] .

4 MULTIMODAL STRESS DETECTION SYSTEM

In this section, we show how we predict stress indicators by leveraging insights from multiple modalities, as shown in Figure 3. Specifically, we analyze a short video clip of a person speaking about a topic on camera and determine if the person is stressed by examining the following four factors: 1) their facial expressions, 2) the prosodic intonation in their voice, 3) the sentiment conveyed in the content of their speech, and 4) any upper body fidgeting. To achieve this, we extract stress signals from each of these modalities using unimodal classifiers and then employ three methods for fusing the predictions of the unimodal classifiers to make a final binary prediction about whether the person is stressed.

4.1 Unimodal Prediction of Stress Indicators

This section discusses how the following individual modalities are used to extract stress indicators from a short video clip with no preprocessing as summarized by Table 1.

4.1.1 Facial Emotion Recognition. This module uses a person's facial expressions to determine stress indicators in a video. To achieve this, we first predict the bounding box containing a person's face for each frame of the video using the MTCNN network [91], which is pre-trained on the FaceNet [72] dataset [5]. The detected bounding boxes are then cropped, and for each crop, we use a Convolutional Neural Network using publicly available weights pre-trained on the Facial Emotion Recognition [26] dataset to predict the probability of expressions - anger, disgust, fear, happiness, sadness, surprise, and neutrality, given a cropped image containing a person's face, which we average across all frames of the clip.

4.1.2 Voice Emotion Recognition. This module aims to identify signs of prosodic stress by analyzing the speaker's voice. To achieve this, we first extract the audio component of each video and pass it through a publicly available Wav2Vec 2.0 model [5], which has been pre-trained on the IEMOCAP dataset [13] using the SpeechBrain toolkit [63] [6]. The model analyzes the prosodic intonation in the person's voice for the entire clip and predicts the probability of them being happy, sad, angry, or neutral.

4.1.3 Sentiment Analysis. In this module, we aim to analyze the sentiment conveyed by the content of the person's speech to infer signs of stress. To do that, we first extract the transcript of the person's speech from the audio component of the video using the Whisper API [62]. We then run inference on a publicly available DistilBERT [68] model pre-trained on the Twitter Sentiment Analysis dataset [69] [7]. The model predicts the probability of the person's sentiment being sadness, joy, love, anger, fear, or surprise from the transcript for the entire clip.

4.1.4 Fidget Detection. Fidgeting is defined as a dynamic self-comforting behavior, as opposed to a static self-comforting behavior [51, 55]. A static self-comforting behavior is a self-comforting behavior where the body parts are not moving, such as a person grabbing their arm with their hand. A dynamic self-comforting behavior is a self-comforting behavior where one of the body parts is moving relative to the other, such as a person rubbing their face with their hand.

[3] https://sites.google.com/view/stress-detection-icmi-24/home
[4] To be compliant with YouTube's copyright policy, we will provide links to the videos sampled along with time stamps of clips, labels for each clip and a script to download the video clips from YouTube automatically

[5] https://github.com/JustinShenk/fer
[6] https://huggingface.co/speechbrain/emotion-recognition-wav2vec2- IEMOCAP/tree/main
[7] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

Modality	Unimodal Stress Prediction Model	Predicted Classes
Face	pre-trained MTCNN [91] for face detection + CNN pre-trained on FER dataset for emotion recognition [26]	anger, disgust, fear, sadness, neutral, happiness, surprise
Voice	Wav2Vec 2.0 model [5] pre-trained on the IEMOCAP dataset [13]	sad, angry, neutral, happy
Sentiment	DistilBERT [68] pre-trained on the Twitter Sentiment Analysis dataset [69]	sadness, anger, fear, joy, love, surprise
Fidgeting	MoveNet [1] for human pose detection and custom fidget detection algorithm	% of video frames with fidgeting

Table 1: Description of each unimodal stress prediction model.

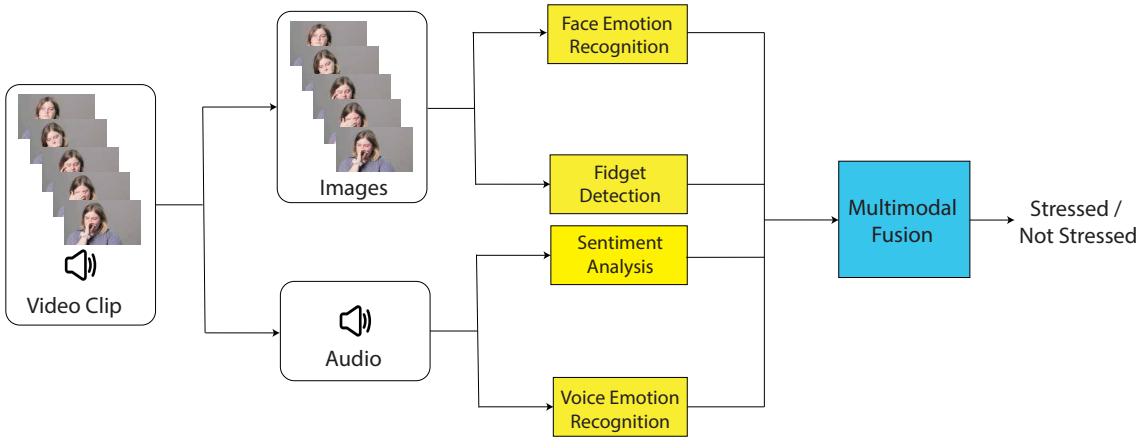

Figure 3: System Diagram. Each video clip is split into a series of images, fed into the facial and fidget stress detectors, and an audio file, fed into the sentiment and voice stress detectors. Then, the output of each of these unimodal nodes passed through our multimodal fusion system to obtain a final prediction.

To identify fidgeting signals that are related to stress from a given video, we first use a pre-trained MoveNet [1] pose detection model that can detect 17 keypoints on a person's body to locate the person's body in the video. Next, we draw polygons around the body parts of interest, like people's hands and faces. Similar to Lin et al. [47], we focus on the movements of the hands relative to other body parts, such as gestures involving touching the face or other hand that are typically associated with fidgeting. Then, we calculate if any polygons overlap to determine if any self-comforting gestures are present. For these overlapping polygons, we employ the Gunnar-Farneback method [21] to calculate the optical flow between the keypoints on successive frames of the video and identify significant movements that are characteristic of fidgeting. We use this information to construct a matrix that tracks the overlap of the person's limbs with every other body part. We then check whether the average optical flow value in each overlap region is above one of several thresholds determined via grid search on videos from the Rhythmic Gestures Corpus [51] (1.5 for hand-on-hand gestures, 0.2 for hand-on-face gestures, and 0.8 for all other gestures) to distinguish between static and dynamic self-comforting behaviors. Then, if over 30% of recent frames contain dynamic self-comforting behaviors, we say that fidgeting is occurring in the current frame.

4.2 Multimodal Fusion

The main objective of multimodal fusion is to take advantage of the unique strengths of every modality to produce more accurate predictions than what can be achieved with a single modality [22]. In this work, we explore three common multimodal fusion techniques, intermediate fusion, voting-based late fusion, and learning-based late fusion, using the predictions of the unimodal classifiers of stress described in Sec. 4.1 in different ways to determine the most effective fusion strategy for predicting stress indicators in videos.

4.2.1 Intermediate Fusion. Intermediate or feature-level fusion combines features or outputs from different modalities before the final classification stage. In our system, we extract embeddings from facial emotion recognition, voice emotion recognition, and sentiment analysis by extracting the output of the last hidden layer of their individual networks. For the fidget detection module, we consider the matrix that tracks the overlap of the person's limbs with every other body part as the feature vector. We then combine these feature vectors into a single vector and train a four-layer Multi-Layer Perceptron (MLP) with hidden sizes [512, 256, 64, 16] (learning rate = 0.001, batch size = 64) for 100 epochs to perform a binary classification task to predict whether a person is stressed. This approach allows the model to learn interactions between different modalities at a feature level, potentially revealing complex patterns contributing to a more nuanced understanding of stress.

4.2.2 Voting-Based Late Fusion. Voting-based late fusion entails aggregating the predictions from each modality using a majority voting scheme. To calculate the vote of the face, voice, and sentiment unimodal node, the voting-based late fusion system will first sum up the values of stress and no-stress predicted classes that the face, voice, and sentiment unimodal predictors, as seen in Table 1. In

other words, we compute the sum of probabilities $X_s = \sum x_s$ for stress indicators, and $X_{ns} = \sum x_{ns}$ for non-stress indicators where x_s and x_{ns} represent the probability score of a stress and non-stress indicator from a given unimodal classifier respectively. For each of these 3 unimodal predictors, if the probability of stress predicted (X_s) is above some empirically determined threshold, a vote is cast in favor of stress. Similarly, if the sum of the probabilities of unimodal non-stress indicators (X_{ns}) is above a different threshold, a vote is cast against stress. Mathematically, for a given unimodal node, we compare X_s and X_{ns} to a stress threshold t_s, and a non-stress threshold t_{ns} respectively, and update the vote count s as follows:

$$s = \begin{cases} s+1 & \text{if } X_s > t_s \\ s-1 & \text{if } X_{ns} > t_{ns} \end{cases}$$

For each modality in face, sentiment and voice, t_s and t_{ns} are calculated as:

$$t_s = \frac{\text{number of classes indicating stress}}{\text{total number of classes}}$$

$$t_{ns} = \frac{\text{number of classes indicating no stress}}{\text{total number of classes}}$$

This is done to consider that there may be different numbers of stressed and non-stressed classes for each modality, so the sum value that constitutes an unimodal node predicting stress over non-stress may vary. This approach also prevents a modality from influencing the final decision of stress or non-stress if a given classifier predicts an equal probability distribution for all classes in a video clip.

The fidget node works slightly differently: if a majority of frames are judged as containing fidgeting, a vote is cast in favor of stress; the fidget node cannot actively vote against stress as the lack of frames containing fidgeting does not necessarily mean that the subject is not stressed. This is because, for many videos in our dataset, a subject's hands may become obscured for several reasons, such as being out of frame or behind another body part. In these instances, our fidget predictor will default to predicting no fidgeting for that frame.

The majority vote across all modalities determines the final stress prediction. If neither sum of votes is above their threshold, no vote is cast by that unimodal node in either direction. For each unimodal predictor, their stress and non-stress indicators, and the threshold values are:

(1) **Face:** The expressions anger, disgust, fear, and sadness are considered indicators of *stress* [18], while the other expressions (happiness, surprise) are considered indicators of *no stress*. The neutral emotional score is not considered to indicate stress or no stress. The threshold for a stress vote (t_s) is 4/7, and the threshold for a no-stress vote (t_{ns}) is 2/7.
(2) **Sentiment:** We classify sadness, anger, and fear as indicators of *stress*, while joy, love, and surprise are considered indicative of *no stress*. The threshold for a stress vote (t_s) is 3/6, and the threshold for a no-stress (t_{ns}) vote is 3/6.
(3) **Voice:** The emotions of sadness and anger are considered indicators of *stress*, while happy is considered an indicator of *no stress*. The neutral emotional score is not considered to indicate stress or no stress. The threshold for a stress vote (t_s) is 2/4, and the threshold for a no-stress vote (t_{ns}) is 1/4.
(4) **Fidgeting:** If more than 50% of frames contain fidgeting behaviors, it suggests that the person is experiencing *stress*.

4.2.3 Learning-Based Late Fusion. Learning-based late fusion or stacking [85] employs a learning model to integrate the final multiclass probability outputs from each unimodal classifier as summarized in the last column of Table 1. Unlike voting-based fusion, where each vote is weighted equally, learning-based fusion allows for learning how each modality's predictions should be weighted to best predict if a person shown in the video clip is stressed or not. For instance, in contexts where facial expressions and prosodic features are particularly telling if a person is stressed, a learning-based model can learn to prioritize these modalities over others. Our system extracts the raw probability scores from each unimodal classifier and concatenates the outputs to form a feature vector. We experiment with training a two layer Multi-Layer Perceptron (MLP) with hidden size of 128 and Random Forest (maximum depth = None, minimum samples a leaf node = 2, minimum samples to split an internal node = 10, number of estimators = 200) separately with this feature vector to learn how the predictions from each unimodal classifier can be combined to predict stress indicators.

5 EXPERIMENTS AND RESULTS

In this work, we first use unimodal classifiers to extract indicators of stress from people's sentiment, facial expressions, vocal prosodic intonation and fidgeting gestures from our MultiAffectStress (MAS) dataset. Then, we compare the performance of intermediate fusion, voting-based late-fusion, and two learning-based late multimodal fusion models on our MAS dataset. Finally, we conduct an ablation study across modalities for each multimodal fusion technique to investigate the contribution of each modality towards the final performance of each of multimodal fusion technique.

As shown in Fig. 5, the F1 scores across the fusion techniques exhibit a clear trend: combinations that integrate more modalities tend to achieve higher F1 scores, highlighting that our selected modalities contain complementary information about predicting if a person is stressed from audio-visual data.

5.1 Comparison between Different Fusion Techniques

We compare the following multimodal fusion techniques to effectively predict whether a person is stressed from short video clips of people from our MultiAffectStress dataset.

5.1.1 Intermediate Fusion. In Section 4.2.1, we explain that we collect feature vectors from four different modules: facial emotion recognition, voice emotion recognition, sentiment analysis, and fidget detection. We use these feature vectors to train a Multi-Layer Perceptron (MLP) model on the training set of our MultiAffectStress dataset. To find the best set of hyperparameters (learning rate = 0.001, batch size = 64, number of epochs=100), we perform a grid search on the validation set of our dataset. Our findings indicate that combining all four modalities leads to an F1 score of 0.73 on the test set. Intermediate fusion produces the lowest F1 score among all the other fusion techniques because when features from multiple

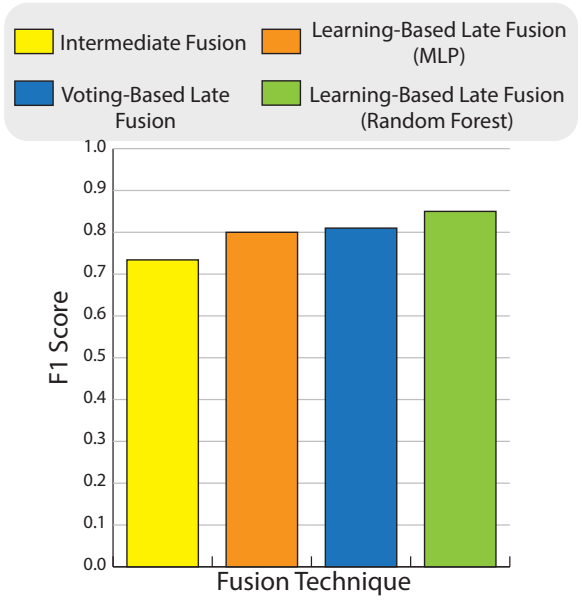

Figure 4: Comparison between different multimodal fusion techniques with four modalities.

modalities are combined, the resulting feature vector is significantly larger than any single-modality feature set. Therefore, models with higher capacity and a larger dataset would be needed to learn complex interactions between the different modalities effectively.

5.1.2 Voting-Based Late Fusion. Section 4.2.2 explains our voting-based late fusion method. We utilize the predictions from each unimodal classifier to identify indicators of stress. Then, we empirically determine the optimal number of modalities and their weightage to predict stress based on the training and validation sets. Our model achieved an F1 score of 0.81 on the test set, outperforming intermediate fusion techniques when using all four modalities of data.

5.1.3 Learning-Based Late Fusion. To combine the final predictions of each unimodal model, we utilize the method described in Section 4.2.3. A feature vector is constructed by concatenating the probability scores from each modality's predictions. We conducted experiments using Multi-Layer Perceptron (MLP) and Random Forest models on the feature vector from the training partition of our dataset. We performed a grid search on the validation set to obtain the best set of hyperparameters for both the MLP (number of epochs = 10, batch size = 64, learning rate = 0.001) and Random Forest (maximum depth = None, minimum samples a leaf node = 2, minimum samples to split an internal node = 10, number of estimators = 200) models. When all four modalities were combined, the Random Forest model achieved a higher F1 score (0.85) than the MLP model (0.80). Overall, the Random Forest method for learning-based late fusion outperformed all other fusion techniques when using all four modalities.

5.2 Ablation Study of Different Modalities

We performed an ablation study to determine the contribution of each modality in predicting stress, using all four multimodal fusion techniques described in Sec. 4.2. First, we evaluate the performance of each modality individually to determine how good each of them is at individually predicting stress. Then, we removed one modality at a time from the input and computed the F1 score of the method in accurately predicting stress in the person. We repeated this process for all four fusion techniques.

5.2.1 Unimodal Analysis of Stress Indicators. Fig 5 a) shows models trained on a single data modality. When compared to Fig. 4 and Fig. 5 a), we observed that unimodal predictors almost always had the lowest performance over using more modalities, indicating that single modalities may struggle to predict stress indicators accurately.

Additionally, we found that the strongest indicator of a person's stress levels is the sentiment expressed in the content of their speech. This can be attributed to the fact that most videos in the MultiAffectStress dataset contain instances of people talking about stressful life experiences. On the other hand, the vocal stress detector appears to perform worse than other modalities. This may be since the Wav2Vec 2.0 model [5] used in the study is trained on an acted dataset, IEMOCAP [13], which may not accurately reflect real-world emotions [31]. Finally, the fidget stress detector performs surprisingly well despite making up the smallest proportion of labels in the dataset (as shown by Fig. 2 b)), indicating that self-comforting gestures can be an important indicator of stress.

5.2.2 Impact of Adding Modalities on Stress Prediction. Fig. 5 b) shows that the addition of more modalities leads to an improvement in performance for stress detection. As sentiment stress has been shown to be the strongest indicator of stress in Fig. 5 a), Fig. 5 c) investigates the impact of adding and removing the sentiment modality on the performance of each multimodal fusion model when three modalities are used.

In Fig. 5 c), the "verbal" bars show the mean and standard deviation of F1 scores when sentiment stress is combined with two of fidget, face, or voice emotion recognition modalities. The "non-verbal" bars indicate the F1 score of our model when sentiment stress is excluded. When sentiment analysis is combined with two of fidget, face, or voice emotion recognition (verbal), all three late fusion techniques significantly improve performance over the non-verbal modality combination (face + voice + fidget). However, we find that intermediate fusion results in higher performance than all combinations where sentiment is included when the sentiment modality is excluded (non-verbal). This suggests that intermediate fusion is particularly effective at leveraging non-verbal cues and extracting nuanced stress indicators by analyzing complex features across modalities for our MultiAffectStress dataset.

6 LIMITATIONS AND FUTURE WORK

This work has several limitations. First, as demonstrated in Fig. 1, we selected video clips where a person was very clearly stressed or not stressed as outlined by our clip selection process in Sec. 3. Even though this helped us achieve a high IRR, it also means that models trained on this dataset may not be able to pick up on more

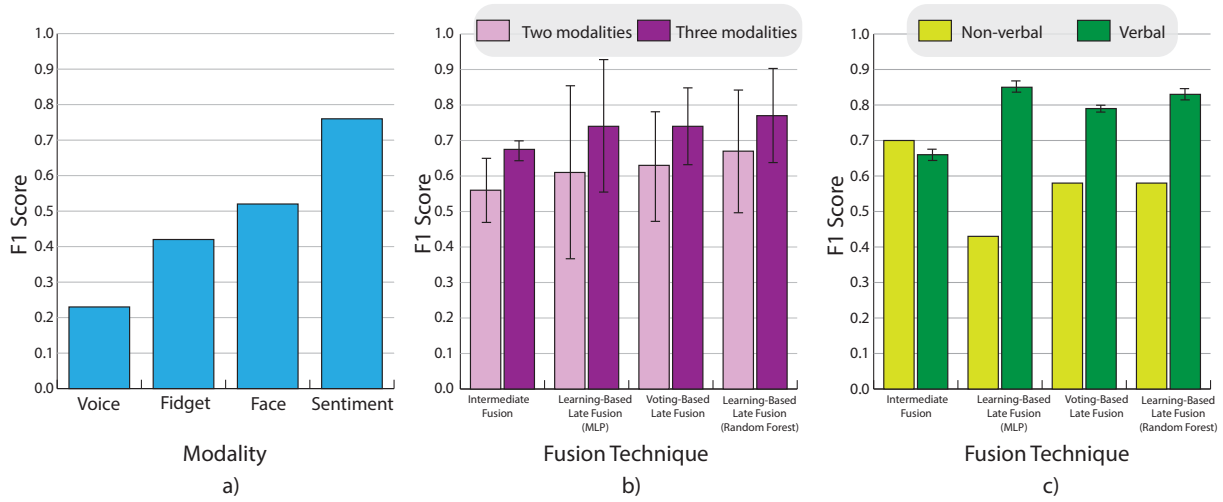

Figure 5: a) Performance of unimodal stress prediction models. b) Comparison between the combination of two vs three modalities for all four multimodal fusion techniques. c) Comparison between fusion of non-verbal stress indicators (facial, vocal, and fidgeting) and verbal stress indicators (mean and standard deviation for combining sentiment with two other modalities between facial, voice, and fidgeting) for all four multimodal fusion techniques.

subtle signs of stress, which could impact how widely applicable they would be. Therefore, an interesting direction to take would be to curate another dataset with less obvious expressions of stress, to train models that can learn subtle stress indicators.

In a similar vein, the videos that we have chosen for our dataset show individuals recalling and discussing past stressful events. However, this is very different from experiencing stress in real-time, as the expression of stress while recalling an event can be quite different from expressions during the actual occurrence of stress. Therefore, further research could benefit from incorporating data sources that capture real-time stress responses to better understand and detect stress in real-world scenarios, where non-verbal cues may be more prevalent or immediate. These narrations also tend to focus heavily on the verbal expression of stress, specifically through sentiment. This can significantly impact the training of our multimodal fusion models, causing them to be biased towards recognizing sentiment stress.

In this work, we only benchmark our model with our binary *stress* or *no stress* labels. Future work could additionally use our unimodal labels to fine-tune the publicly available emotion recognition classifiers before training the multimodal fusion models. This would potentially enhance the identification of contributing factors to stress levels in an individual.

Additionally, developing a model that performs well on a dataset differs from deploying it in real-world settings, such as on social robots [3, 53, 67, 71]. The data in the real world may be far noisier, with multiple speakers or people in a frame, background noise, partially obscured subjects, poor lighting, hardware differences, and more varied and subtle expressions of stress. Therefore, fine-tuning stress detection models with real-world data for continual stress monitoring in the wild is left for future work.

Finally, it is also worth considering the ethical implications of our stress detection system. As our system is purely audio-visual, utilizing publicly displayed signals, there are not the same privacy concerns associated with biometric stress detectors. Instead, the focus should be on the ways our system could be deployed. While it may be harmless when used by a shopping mall robot, the same cannot necessarily be said when deploying it for use in airport security checkpoints, or when screening job applicants.

7 CONCLUSION

In this work, we found that combining different modalities such as facial expressions, vocal prosody, sentiment analysis, and physical fidgeting using various multimodal fusion techniques can capture a comprehensive spectrum of stress indicators. Due to the lack of available datasets with these modalities, we sampled videos from YouTube and annotated for stress-related indicators across these four modalities. Our experimental results support our hypothesis that a multimodal approach is more effective than unimodal methods. Our models achieved a maximum F1 score of 0.85, demonstrating the potential of combining different modal inputs to improve stress detection systems. Ablation studies provided insights into how each modality contributes to the detection process, and we found that the combined use of all modalities consistently resulted in the highest performance metrics across all tested multimodal fusion techniques.

ACKNOWLEDGMENTS

The authors would like to thank Ellie Mamantov, Shasvat Desai, Drazen Brzcic, and Anant Srinivasan for their support and feedback in improving the paper. This work was funded by the National Science Foundation (NSF) under grants No. 1955653 and 2106690 and the National Institutes of Health (NIH) under grant No. R44MD017104. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or NIH.

REFERENCES

[1] [n. d.]. Next-Generation Pose Detection with MoveNet and TensorFlow.js. https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html

[2] Muhammad Abdullah, Mobeen Ahmad, and Dongil Han. 2021. Hierarchical attention approach in multimodal emotion recognition for human robot interaction. In *2021 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. IEEE, 1–4.

[3] Timothy Adamson, Debasmita Ghose, Shannon C Yasuda, Lucas Jehu Silva Shepard, Michal A Lewkowicz, Joyce Duan, and Brian Scassellati. 2021. Why we should build robots that both teach and learn. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 187–196.

[4] Fares Al-Shargie, Masashi Kiguchi, Nasreen Badruddin, Sarat C Dass, Ahmad Fadzil Mohammad Hani, and Tong Boon Tang. 2016. Mental stress assessment using simultaneous measurement of EEG and fNIRS. *Biomedical optics express* 7, 10 (2016), 3882–3898.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[6] Serdar Baltaci and Didem Gokcay. 2016. Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features. *International Journal of Human–Computer Interaction* 32, 12 (2016), 956–966.

[7] Tanja Bänziger and Klaus R Scherer. 2010. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook* 2010 (2010), 271–94.

[8] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Sequeira, Alexander Sutherland, and Stefan Wermter. 2018. The OMG-Emotion Behavior Dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)* (Rio de Janeiro, Brazil). IEEE, 1408–1414. https://doi.org/10.1109/IJCNN.2018.8489099

[9] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J. Russell, and Michael Wong. 2004. "You Stupid Tin Box" - Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus. In *International Conference on Language Resources and Evaluation*. https://api.semanticscholar.org/CorpusID:1027542

[10] Linda Becker, Alexander Heimerl, and Elisabeth André. 2023. ForDigitStress: presentation and evaluation of a new laboratory stressor using a digital job interview-scenario. *Frontiers in Psychology* 14 (2023). https://doi.org/10.3389/fpsyg.2023.1182959

[11] Laura Boccanfuso, Quan Wang, Iolanda Leite, Beibin Li, Colette Torres, Lisa Chen, Nicole Salomons, Claire Foster, Erin Barney, Yeojin Amy Ahn, et al. 2016. A thermal emotion classifier for improved human-robot interaction. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 718–723.

[12] Margaret M Bradley and Peter J Lang. 2000. Measuring emotion: Behavior, feeling, and physiology. (2000).

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42 (2008), 335–359.

[14] Sara Campanella, Ayham Altaleb, Alberto Belli, Paola Pierleoni, and Lorenzo Palma. 2023. A method for stress detection using empatica E4 bracelet and machine-learning techniques. *Sensors* 23, 7 (2023), 3565.

[15] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing* 5, 4 (2014), 377–390.

[16] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. 2023. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision* 131, 6 (2023), 1346–1366.

[17] Juan Abdon Miranda Correa, Mojtaba Khomami Abadi, Niculae Sebe, and I. Patras. 2017. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* 12 (2017), 479–493. https://api.semanticscholar.org/CorpusID:8743034

[18] Lynn A. Fairbanks, Michael T. McGuire, and Candace J. Harris. 1982. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of Abnormal Psychology* 91, 2 (1982), 109–119. https://doi.org/10.1037/0021-843X.91.2.109

[19] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. 2010. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia* 12 (2010), 591–598. https://api.semanticscholar.org/CorpusID:5326393

[20] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool. 2010. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia* 12, 6 (2010), 591–598.

[21] Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*. Springer, 363–370.

[22] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32, 5 (2020), 829–864.

[23] Mihai Gavrilescu and Nicolae Vizireanu. 2019. Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors* 19, 17 (2019), 3693.

[24] Debasmita Ghose, Shasvat M Desai, Sneha Bhattacharya, Deep Chakraborty, Madalina Fiterau, and Tauhidur Rahman. 2019. Pedestrian detection in thermal images using saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[25] Giorgos Giannakakis, Dimitris Manousos, Vaggelis Chaniotakis, and Manolis Tsiknakis. 2018. Evaluation of head pose features for stress detection and classification. In *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*. 406–409. https://doi.org/10.1109/BHI.2018.8333454

[26] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III 20*. Springer, 117–124.

[27] Rain Eric Haamer, Eka Rusadze, Iiris Lüsi, Tauseef Ahmed, Sergio Escalera, and Gholamreza Anbarjafari. 2017. Review on Emotion Recognition Databases. In *Human-Robot Interaction*, Gholamreza Anbarjafari and Sergio Escalera (Eds.). IntechOpen, Rijeka, Chapter 3. https://doi.org/10.5772/intechopen.72748

[28] Yu He, Licai Sun, Zheng Lian, Bin Liu, Jianhua Tao, Meng Wang, and Yuan Cheng. 2022. Multimodal Temporal Attention in Sentiment Analysis. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*. 61–66.

[29] Muhammad Syazani Hafiy Hilmy, Ani Liza Asnawi, Ahmad Zamani Jusoh, Khaizuran Abdullah, Siti Noorjannah Ibrahim, Huda Adibah Mohd Ramli, and Nor Fadhillah Mohamed Azmin. 2021. Stress classification based on speech analysis of MFCC feature via machine learning. In *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 339–343.

[30] Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2020. MuSE: a Multimodal Dataset of Stressed Emotion. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 1499–1510. https://aclanthology.org/2020.lrec-1.187

[31] Rebecca Jürgens, Annika Grass, Matthis Drolet, and Julia Fischer. 2015. Effect of Acting Experience on Emotion Expression and Recognition in Voice: Non-Actors Provide Better Stimuli than Expected. *Journal of Nonverbal Behavior* 39, 3 (01 Sep 2015), 195–214. https://doi.org/10.1007/s10919-015-0209-5

[32] C Jyotsna, J Amudha, Amritanshu Ram, and Giandomenico Nollo. 2023. IntelEye: An intelligent tool for the detection of stressful state based on eye gaze data while watching video. *Procedia Computer Science* 218 (2023), 1270–1279.

[33] Manasa Kalanadhabhatta, Shaily Roy, Trevor Grant, Asif Salekin, Tauhidur Rahman, and Dessa Bergen-Cico. 2023. Detecting PTSD Using Neural and Physiological Signals: Recommendations from a Pilot Study. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[34] Kyriaki Kalimeri and Charalampos Saitis. 2016. Exploring multimodal biosignal features for stress detection during indoor mobility. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 53–60. https://doi.org/10.1145/2993148.2993159

[35] Mitchel Kappen, Kristof Hoorelbeke, Nilesh Madhu, Kris Demuynck, and Marie-Anne Vanderhasselt. 2022. Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods* (2022), 1–12.

[36] N Keshan, PV Parimi, and Isabelle Bichindaritz. 2015. Machine learning for stress detection from ECG signals in automobile drivers. In *2015 IEEE International conference on big data (Big Data)*. IEEE, 2661–2669.

[37] Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin G. McInnis, and Emily Mower Provost. 2018. The PRIORI Emotion Dataset: Linking Mood to Emotion Detected In-the-Wild. *ArXiv* abs/1806.10658 (2018). https://api.semanticscholar.org/CorpusID:49523812

[38] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. 2018. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry investigation* 15, 3 (2018), 235.

[39] Dimitrios Kollias and Stefanos Zafeiriou. 2019. Aff-Wild2: Extending the Aff-Wild Database for Affect Recognition. arXiv:1811.07770 [cs.CV]

[40] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Björn Schuller, Kam Star, Elnar Hajiyev, and Maja Pantic. 2019. SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2019), 1022–1040. https://api.semanticscholar.org/CorpusID:57759395

[41] Satish Kumar, ASM Iftekhar, Michael Goebel, Tom Bullock, Mary H MacLean, Michael B Miller, Tyler Santander, Barry Giesbrecht, Scott T Grafton, and BS Manjunath. 2021. StressNet: detecting stress in thermal videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 999–1009.

[42] Iulia Lefter, Gertjan J Burghouts, and Leon JM Rothkrantz. 2014. An audio-visual dataset of human–human interactions in stressful situations. *Journal on Multimodal User Interfaces* 8 (2014), 29–41.

[43] Iulia Lefter, Gertjan J. Burghouts, and Léon J. M. Rothkrantz. 2014. An audio-visual dataset of human–human interactions in stressful situations. *Journal on Multimodal User Interfaces* 8 (2014), 29–41. https://api.semanticscholar.org/CorpusID:207402069

[44] Wei Li, Farnaz Abtahi, Christina Tsangouri, and Zhigang Zhu. 2016. Towards an "In-the-Wild" Emotion Dataset Using a Game-Based Framework. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2016), 1526–1534. https://api.semanticscholar.org/CorpusID:4629457

[45] Yuan Li, Du Li, Yiyu Xu, Xuelin Yuan, and Xiangwei Zhu. 2024. Human State Recognition Using Ultra Wideband Radar Based On CvT. *IEEE Internet of Things Journal* (2024).

[46] Chung-Yen Liao, Rung-Ching Chen, and Shao-Kuo Tai. 2018. Emotion stress detection using EEG signal and deep learning technologies. In *2018 IEEE International Conference on Applied System Invention (ICASI)*. IEEE, 90–93.

[47] Weizhe Lin, Indigo Orton, Mingyu Liu, and Marwa Mahmoud. 2020. Automatic Detection of Self-Adaptors for Psychological Distress. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. 371–378. https://doi.org/10.1109/FG47880.2020.00032

[48] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[49] Yunfei Luo, Iman Deznabi, Abhinav Shaw, Natcha Simsiri, Tauhidur Rahman, and Madalina Fiterau. 2024. Dynamic clustering via branched deep learning enhances personalization of stress prediction from mobile sensor data. *Scientific Reports* 14, 1 (2024), 6631.

[50] Marwa Mahmoud, Louis-Philippe Morency, and Peter Robinson. 2013. Automatic multimodal descriptors of rhythmic body movement. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. 429–436.

[51] Marwa Mahmoud, Louis-Philippe Morency, and Peter Robinson. 2013. Automatic Multimodal Descriptors of Rhythmic Body Movement. In *International Conference on Multimodal Interaction*.

[52] Kayla Matheus, Ellie Mamantov, Marynel Vázquez, and Brian Scassellati. 2023. Deep Breathing Phase Classification with a Social Robot for Mental Health. In *Proceedings of the 25th International Conference on Multimodal Interaction*. 153–162.

[53] Kayla Matheus, Marynel Vázquez, and Brian Scassellati. 2022. A social robot for anxiety reduction via deep breathing. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 89–94.

[54] Naoshi Matsuo, Nobuyuki Washio, Shouji Harada, Akira Kamano, Shoji Hayakawa, and Kazuya Takeda. 2011. A study of psychological stress detection based on the non-verbal information. *IEICE Technical Report; IEICE Tech. Rep.* 111, 97 (2011), 29–33.

[55] Albert Mehrabian and Shan L Friedman. 1986. An analysis of fidgeting and associated individual differences. *Journal of Personality* 54, 2 (1986), 406–429. https://doi.org/10.1111/j.1467-6494.1986.tb00402.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-6494.1986.tb00402.x

[56] Scott M. Monroe. 2008. Modern Approaches to Conceptualizing and Measuring Human Life Stress. *Annual Review of Clinical Psychology* 4, Volume 4, 2008 (2008), 33–52. https://doi.org/10.1146/annurev.clinpsy.4.022007.141207

[57] Tanya Nijhawan, Girija Attigeri, and T Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data* 9, 1 (2022), 33.

[58] Rosalind W Picard. 2016. Automating the recognition of stress and emotion: From lab to real-world impact. *IEEE MultiMedia* 23, 3 (2016), 3–7.

[59] Anthony J Porcelli and Mauricio R Delgado. 2017. Stress and decision making: effects on valuation, learning, and risk-taking. *Current Opinion in Behavioral Sciences* 14 (2017), 33–39. https://doi.org/10.1016/j.cobeha.2016.11.015 Stress and behavior.

[60] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information fusion* 37 (2017), 98–125.

[61] Colin Puri, Leslie Olson, Ioannis Pavlidis, James Levine, and Justin Starren. 2005. StressCam: non-contact measurement of users' emotional states through thermal imaging. In *CHI'05 extended abstracts on Human factors in computing systems*. 1725–1728.

[62] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.

[63] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. 2021. SpeechBrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624* (2021).

[64] Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (2013), 1–8. https://api.semanticscholar.org/CorpusID:206651806

[65] Sunita Sahu, Ekta Kithani, Manav Motwani, Sahil Motwani, and Aadarsh Ahuja. 2021. Stress Detection of Office Employees Using Sentiment Analysis. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2*. Springer, 143–153.

[66] Lizawati Salahuddin and Desok Kim. 2006. Detection of acute stress by heart rate variability using a prototype mobile ECG sensor. In *2006 International Conference on Hybrid Information Technology*, Vol. 2. IEEE Computer Society, 453–459.

[67] Nicole Salomons, Tom Wallenstein, Debasmita Ghose, and Brian Scassellati. 2022. The Impact of an In-Home Co-Located Robotic Coach in Helping People Make Fewer Exercise Mistakes. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 149–154.

[68] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[69] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3687–3697. https://doi.org/10.18653/v1/D18-1404

[70] Pritam Sarkar, A L Posen, and Ali Etemad. 2022. AVCAffe: A Large Scale Audio-Visual Dataset of Cognitive Load and Affect for Remote Work. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:248811751

[71] Brian Scassellati, Laura Boccanfuso, Chien-Ming Huang, Marilena Mademtzi, Meiying Qin, Nicole Salomons, Pamela Ventola, and Frederick Shic. 2018. Improving social skills in children with ASD using a long-term, in-home social robot. *Science Robotics* 3, 21 (2018), eaat7544.

[72] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[73] Hashini Senaratne, Kisrten Ellis, Sharon Oviatt, and Glenn Melvin. 2020. Detecting and differentiating leg bouncing behaviour from everyday movements using tri-axial accelerometer data. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 127–130.

[74] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. 2009. Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine* 14, 2 (2009), 410–417.

[75] Abhinav Shaw, Natcha Simsiri, Iman Deznaby, Madalina Fiterau, and Tauhidur Rahaman. 2019. Personalized student stress prediction with deep multitask network. *arXiv preprint arXiv:1906.11356* (2019).

[76] Jonghoon Shin, Junhyung Moon, Beomsik Kim, Jihwan Eom, Noseong Park, and Kyoungwoo Lee. 2021. Attention-based stress detection exploiting non-contact monitoring of movement patterns with IR-UWB radar. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 637–640.

[77] Reza Arefi Shirvan, Seyed Kamaledin Setaredan, and Ali Motie Nasrabadi. 2018. Classification of mental stress levels by analyzing fNIRS signal using linear and non-linear features. *International Clinical Neuroscience Journal* 5, 2 (2018), 55.

[78] M. Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3 (2012), 42–55. https://api.semanticscholar.org/CorpusID:2820480

[79] Savita Sondhi, Munna Khan, Ritu Vijay, Ashok K Salhan, et al. 2015. Vocal indicators of emotional stress. *International Journal of Computer Applications* 122, 15 (2015), 38–43.

[80] Lukas Stappen, Alice Baird, Lea Schumann, and Björn W. Schuller. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. *IEEE Transactions on Affective Computing* 14 (2021), 1334–1350. https://api.semanticscholar.org/CorpusID:231627534

[81] Nattapong Thammasan, Koichi Moriyama, Ken-ichi Fukui, and Masayuki Numao. 2017. Familiarity effects in EEG-based emotion recognition. *Brain informatics* 4 (2017), 39–50.

[82] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label studio: Data labeling software. *Open source software available from https://github. com/heartexlabs/label-studio* 2022 (2020).

[83] Andrea Vidal, Ali N. Salman, Wei-Cheng Lin, and Carlos Busso. 2020. MSP-Face Corpus: A Natural Audiovisual Emotional Database. *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020). https://api.semanticscholar.org/CorpusID:224816670

[84] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Seattle, Washington) *(UbiComp '14)*. Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/2632048.2632054

[85] David H Wolpert. 1992. Stacked generalization. *Neural networks* 5, 2 (1992), 241–259.

[86] Kieran Woodward and Eiman Kanjo. 2020. iFidgetCube: Tangible Fidgeting Interfaces (TFIs) to Monitor and Improve Mental Wellbeing. *IEEE Sensors Journal* 21, 13 (2020), 14300–14307.

[87] Yi Xiao, Felipe Codevilla, Akhil Gurram, Onay Urfalioglu, and Antonio M López. 2020. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* 23, 1 (2020), 537–547.

[88] Yi Xiao, Harshit Sharma, Zhongyang Zhang, Dessa Bergen-Cico, Tauhidur Rahman, and Asif Salekin. 2024. Reading Between the Heat: Co-Teaching Body Thermal Signatures for Non-intrusive Stress Detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–30.

[89] Yiqun Yao, Michalis Papakostas, Mihai Burzo, Mohamed Abouelenien, and Rada Mihalcea. 2021. MUSER: MUltimodal Stress Detection using Emotion Recognition as an Auxiliary Task. *CoRR* abs/2105.08146 (2021). arXiv:2105.08146 https://arxiv.org/abs/2105.08146

[90] Jin Zhang, Xue Mei, Huan Liu, Shenqiang Yuan, and Tiancheng Qian. 2019. Detecting negative emotional stress based on facial expression in real time. In *2019 IEEE 4th international conference on signal and image processing (ICSIP)*. IEEE, 430–434.

[91] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[92] Zheng Zhang, Jeffrey M. Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Aybars Ciftci, Shaun J. Canavan, Michael J. Reale, Andy Horowitz, Huiyuan Yang, Jeffrey F. Cohn, Qiang Ji, and Lijun Yin. 2016. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 3438–3446. https://api.semanticscholar.org/CorpusID:6578368