Extracting Multimodal Embeddings via Supervised Contrastive Learning for Psychological Screening

Manasa Kalanadhabhatta
University of Massachusetts Amherst
Amherst, MA, USA
manasak@cs.umass.edu

Adrelys Mateo Santana
University of Massachusetts Amherst
Amherst, MA, USA
amateosantan@umass.edu

Deepak Ganesan
University of Massachusetts Amherst
Amherst, MA, USA
dganesan@cs.umass.edu

Tauhidur Rahman
University of Massachusetts Amherst
Amherst, MA, USA
trahman@cs.umass.edu

Adam Grabell
University of Massachusetts Amherst
Amherst, MA, USA
agrabell@umass.edu

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII) | 978-1-6654-5908-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ACII55700.2022.9953836

***Abstract*—The diagnosis of psychological disorders in early childhood is of utmost importance given their severe impact on children's academic and social skills as well as general adaptive functioning. Wearable and video-based systems have the potential to collect important diagnostic information in the form of neurophysiological and behavioral signals. However, accurate prediction of psychological disorder status from multimodal data streams necessitates their combination into meaningful features for classification models. In this work, we present a multitask supervised contrastive learning approach to learn useful multimodal embeddings from functional Near-Infrared Spectroscopy, galvanic skin response, and facial video data collected during a frustration-inducing task. The generated embeddings are able to accurately infer emotion regulation-related psychological disorders with an F1 score of 0.91, having significant implications for early-childhood mental health diagnoses.**

***Index Terms*—emotion regulation, multimodal representation learning, supervised contrastive learning**

I. Introduction

Learning to modulate the duration, valence, or intensity of an emotional experience, or emotion regulation, is central to healthy development in early childhood [1]. Poor emotion regulation, especially in response to negative emotional challenges, has been linked to several psychological disorders such as attention deficit hyperactivity disorder (ADHD; [2]), conduct problems [3], and childhood depression [4]. Emotion dysregulation can also cause impairments in academic [5], social [6], and adaptive functioning [1]. However, it is often difficult to distinguish normative misbehavior in early childhood from dysregulation-related psychological disorders [7]. Current methods to detect early psychopathology are difficult to access, need to be administered by trained professionals, and have modest diagnostic accuracy [8].

Wearables and video-based tools have long been utilized in affective computing research to detect emotion [9] as well as to assess emotion regulation [10]. However, their utility in screening for emotion dysregulation-related mental disorders, especially in early childhood, has been severely under-explored. Research in clinical psychology has identified several promising behavioral and neurophysiological signals that are indicative of psychopathology. For instance, decreased neural activation in the lateral prefrontal cortex (PFC) during emotion regulation has been linked to higher aggressive behavior [3] and ADHD [11]. Recent advances in neuroimaging have made it easier to measure lateral PFC activation in younger populations using non-invasive techniques such as functional Near-Infrared Spectroscopy (fNIRS). Dysfunctional emotion-related lateral PFC-amygdala connectivity has also been associated with conduct disorders [12], and galvanic skin response (GSR) has been identified as an easily-observable byproduct of amygdala activation [13]. Facial expressions, which can be observed via video, have additionally been found to be strong behavioral correlates of lateral PFC activation and emotion regulation in children [14].

This work attempts to utilize PFC neural activation, GSR, and facial videos recorded during a clinically validated emotion regulation task [15] to classify preschool-aged children with and without psychopathological symptoms. We first establish the feasibility of detecting psychological disorder status using research-informed handcrafted features from each of these modalities. Following this, we investigate whether it is possible to combine information from all three sources in an effective manner to improve classification performance. To this end, we extract multimodal embedding features using a multitask supervised contrastive learning approach.

Our approach addresses two related challenges in multimodal learning in the affective computing domain. First, handcrafted feature extraction from behavioral or neurophysiological data may not always be possible due to several reasons (e.g., missing data, short observation durations, ephemeral events of interest, etc.), limiting the amount of usable data to train machine learning models. Second, combining several sources of data effectively in order to learn cross-modal features is still a challenge, though recent efforts such as [16]

This work was supported by the National Institute of Mental Health grant NIMH K23 MH111708 and the National Science Foundation grants 1839999, 1951928, and 1815347.

and [17] have made some progress in this direction. We overcome these issues by proposing a novel approach for learning multimodal embeddings using supervisory signals from domain-specific labels and task characteristics. This allows us to leverage more of the available data and simultaneously learn cross-modality features for model training.

Our proposed method successfully identifies children with psychopathology with an F1 score of 0.91 and area under the receiver operating characteristic curve (AUROC) of 0.9. This is a significant improvement over the predictive performance of baseline methods as well as current clinical diagnostic tools [8], with important implications for technology-enabled mental health screening.

II. Related Work

A. Learning Multimodal Representations

Several approaches have been explored for learning multimodal representations for affective computing applications. For instance, Mai et al. [17] performed sequential local and global fusion of language, visual, and acoustic data by aligning feature vectors across modalities, while Gu et al. [18] proposed a similar hierarchical fusion of text and audio data for emotion recognition. Hazarika et al. [16] projected representations of each modality onto modality-invariant and -specific subspaces and fused them to obtain sentiment predictions. Other works have used feature-based fusion strategies combining handcrafted features from each modality [19], trained networks with modality-specific heads to deal with missing data [20], or used one modality as a supervisory signal for another to learn representations [21].

However, multimodal representation learning using video and neurophysiological data such as fNIRS or GSR has been investigated in less detail. Tan et al. [22] performed emotion recognition using facial expressions and electroencephalography (EEG) by training modality-specific models and combining their predictions. Torres et al. [19] attempted to combine mean face shape from videos with features from EEG and GSR data for emotion recognition, but did not find face shape to not be an informative feature. This underscores the limitations of handcrafted feature combination, which is also observed in our baseline models. Multimodal emotion recognition work on the AMIGOS dataset [23] has also explored input modalities of video, EEG and GSR. However, EEG is less localized than fNIRS in measuring neural activation, and is highly susceptible to motion artifacts, limiting its use in younger participants.

B. Contrastive and Multitask Learning

Recent work in representation learning has increasingly focused on self-supervised learning approaches to leverage unlabeled data during training. One popular approach for learning visual representations is SimCLR [24], which proposed a contrastive learning framework that minimizes the distance between representations of different augmentations of the same image. Contrastive learning has been used for affective computing tasks such as emotion recognition from speech [25] and action unit detection from facial images [26].

Khosla et al. [27] extended the idea of contrastive learning to the supervised setting, where additional information about the training data is available in the form of classification labels. In this case, the contrastive loss is updated to minimize the cosine distance between representations of each pair of positive samples in a training batch. Supervised contrastive learning has been used to learn visual [28], audio [29], or language [30] representations, but remains largely unexplored for learning multimodal behavioral and neural data representations.

Prior work has also utilized the multitask learning paradigm, with auxiliary tasks learned simultaneously with the main task, to regularize models. This has been shown to improve predictive performance in various domains, including affect recognition [31] and stress detection [32]. Recent approaches have also explored incorporating weakly-supervised labels [33] or data attributes [34] as auxiliary information into the contrastive learning framework to learn better visual representations. We propose similarly augmenting the supervised contrastive learning approach using labels based on the structure of the task completed by participants in our study.

III. Methodology

A. Emotion Regulation Task

We conducted a study with 94 participants aged 3.5 to 5 years old who completed a clinically validated emotion regulation task in a laboratory setting (see [35] for more details). None of the participants had any psychotic symptoms, existing diagnoses of developmental or intellectual disabilities, or history of head trauma with loss of consciousness. The study was approved by the Institutional Review Board of the University of Massachusetts Amherst.

Participants completed a frustration-inducing task called Incredible Cake Kids [15], where they were asked to select the "most delicious" cake for customers of a virtual bakery. The task consisted of 30 trials during which a virtual customer appeared on the screen along with three virtual cakes for four seconds, followed by two seconds in which the child touched the cake they thought was the most delicious, two seconds of anticipation, and two seconds of positive (e.g., happy) or negative (e.g., grumpy) feedback provided by the virtual customer. Unknown to the child, the feedback provided in each trial was predetermined and was organized into three negative (four negative and one positive trial grouped together) and three positive blocks (four positive and one negative trial), separated by 20-second rest periods between blocks.

B. Psychopathology Assessment

The participants' caregivers also completed various clinical questionnaires designed to measure severity of the most commonly diagnosed early psychological disorders. The children's frequency of ADHD symptoms was recorded using the *ADHD Rating Scale-5 Home Version* [36] and scored on two subscales – ADHD Inattention and ADHD Hyperactivity. The temper loss subscale from the *Multidimensional Assessment Profile for Disruptive Behavior* [37] was used as a measure of child irritability, a transdiagnostic symptom observed in over a

dozen DSM-5 disorders. Caregivers also reported their child's behaviors via the *Child Behavior Checklist* [38] for ages 1.5 to 5, and the subscale on externalizing symptoms was scored separately to measure behaviors consistent with aggression, defiance, non-compliance, and rule-breaking. Participants were categorized as *clinical* (i.e., exhibiting symptoms of psychopathology) if they scored above clinical threshold on at least one of the four subscales, and *non-clinical* otherwise. In this work, we attempt to differentiate between clinical and non-clinical participants using behavioral and neurophysiological signals recorded during the emotion regulation task.

C. Behavioral and Neurophysiological Data

Three sources of behavioral and neurophysiological signals were recorded during the emotion regulation task. These included (i) facial expressions using video cameras, (ii) neural activation in the prefrontal cortex via fNIRS, and (iii) arousal of the autonomic nervous system via GSR.

The complete duration of the task was video recorded using a commercial high definition camera (Axis Communications PTZ Network Camera) pointed at the child's face. Video was recorded at a resolution of 1080p and frame rate of 60 Hz. Participants' neural activity was measured using a NIRx NIRScout imaging system, with an fNIRS probe consisting of eight light-source emitters with 760nm and 850nm LED lights and four detectors. The sources and detectors were attached to an elastic cap with an average inter-optode distance of 3 cm, resulting in ten channels extending over Brodmann areas 10 (ventrolateral prefrontal cortex) and 46 (dorsolateral prefrontal cortex) on both the left and right hemisphere. Raw intensity at each channel was recorded at 7.81 Hz and downsampled to 4 Hz for further analysis. GSR data was collected at a 1000 Hz sampling rate using a MindWare 8-slot BioNex Chassis with disposable foam electrodes applied to the child's non-dominant hand to reduce motion artifacts.

IV. Baseline Models with Explainable Features

We first attempted to classify participants with and without psychological disorders using explainable features extracted from each sensing modality. We leveraged the multiple trials of the frustration-inducing task described in Section III-A by extracting trial-level features from data collected within six seconds of positive or negative feedback for each trial where a participant selected a cake. These trial-level features were used to train a gradient boosting classifier to predict whether the observed trial was from a clinical or non-clinical individual. The predicted probabilities were averaged across all trials to obtain an individual-level prediction.

A. Feature Extraction

The GSR data was processed using the Neurokit2 library, and was first high-pass filtered using a fourth order Butterworth filter with a cutoff frequency of 3 Hz. The filtered signal was then decomposed into tonic and phasic components and divided into epochs of six second durations following positive or negative feedback. We then extracted the maximum amplitude of the phasic component within the epoch, the amplitude of the Skin Conductance Response (SCR) following the feedback, as well as the rise time and recovery time of the SCR signal. Handcrafted SCR features could only be obtained from trials where a significant SCR peak could be observed in the GSR signal – trials with no detected peaks were not considered.

Facial videos recorded during the frustration-inducing task were used to extract anatomical muscle movements using the Facial Action Coding System (FACS). We used OpenFace 2.0 to detect the presence of 18 action units (AUs) within each frame. These include AU01 (inner brow raiser), AU02 (outer brow raiser), AU04 (brow lowerer), AU05 (upper lid raiser), AU06 (cheek raiser), AU07 (lid tightener), AU09 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips part), AU26 (jaw drop), AU28 (lip suck), and AU45 (blink). Trials where OpenFace was unable to successfully track any frames were discarded.

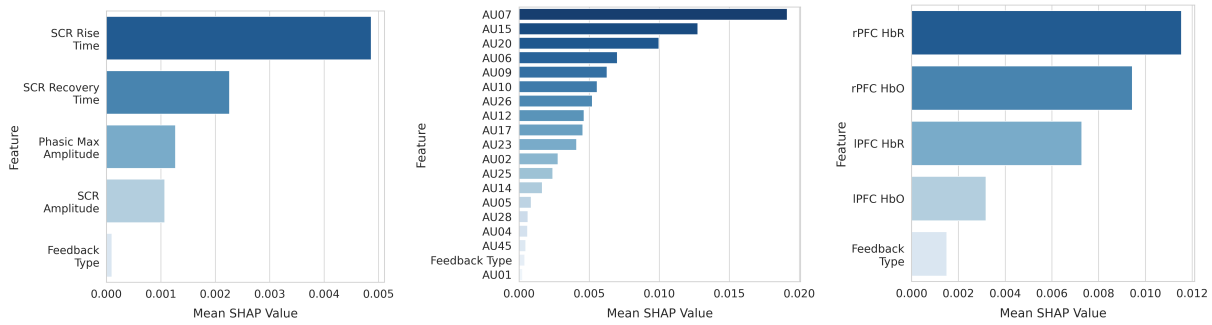
We used the MNE-NIRS library to process the fNIRS data by downsampling it to 4 Hz and converting the raw intensity at each channel to change in optical density (ΔOD). This was further converted to changes in oxyhemoglobin and deoxyhemoglobin (ΔHbO and ΔHbR, respectively) using the modified Beer-Lambert Law. The neural activation level at each trial was then extracted by fitting a generalized linear model of the canonical hemodynamic response function using a first-level design matrix with a cosine drift model. The HbO and HbR activation levels were aggregated across channels into two regions of interest – the left and right prefrontal cortex (lPFC and rPFC) – leading to four activation features for use in our classification models.

In addition to modality-specific features, we also added one feature to each set indicating whether the trial was a positive or negative feedback trial.

B. Training and Evaluation

We used the above-mentioned features to train and evaluate both modality-specific and multimodal psychopathology prediction models. To this end, we partitioned our data into train and test sets with non-overlapping users stratified by their clinical disorder status. The training and test sets contained 60 and 16 participants respectively. Models were trained on all trials within the training set from which handcrafted features could be extracted for a particular modality, or, in case of the multimodal model, for all modalities. The number of valid trials used for training each model are reported in Table I.

We used 3-fold cross validation stratified by psychopathology labels for hyperparameter tuning and model selection. We chose gradient boosting trees as the classifier since it is fairly robust to overfitting and is explainable. All features were standardized before passing them to the classifier. The hyperparameters tuned included the number of features selected (all, top 5 based on mutual information), maximum depth of individual regression estimators (2, 3, 5, 10) and the



(a) GSR features (b) Video features (c) fNIRS features

Fig. 1: Interpreting modality-specific psychopathology classification models using mean SHAP values for handcrafted features.

number of estimators (5, 10, 20, 30). The best performing model based on trial-level F1 scores during cross validation was evaluated on the held out test set. The individual-level prediction performances on the test set are reported in Table I. We observe that the gradient boosting classifier using GSR features outperformed all others, achieving an AUROC of 0.78 and F1 score of 0.67. The classifier using features from all modalities had the lowest AUROC and F1 scores of 0.62 and 0.57 respectively.

TABLE I: Prediction performance using handcrafted features.

Modality	**Number of Features**	**Number of Valid Trials**	**Prediction Performance**	
			AUROC	***F1 Score***
GSR	5	764	0.78	0.67
Video	19	1483	0.64	0.67
fNIRS	5	1498	0.74	0.67
All	27	696	0.62	0.57

Figure 1 shows the importance of each feature used for prediction in the modality-specific classifiers in terms of its mean SHapley Additive exPlanation (SHAP; [39]) values. We find that the SCR rise and recovery times contributed most to predictions in the classifier trained on GSR features. Among video features, AU07 is the most informative while rPFC features dominated the predictions for the fNIRS model.

We also observe that the classifier trained on features from all modalities failed to achieve comparable prediction performance to the modality-specific classifiers, demonstrating that feature combination through concatenation may not add useful information (similar to [19]). This issue is compounded by the availability of fewer trials (696) where handcrafted features could be extracted from all modalities.

C. Limitations of Baseline Models

While the modality-specific models discussed above achieve reasonable prediction performance and have the advantage of being interpretable, it is clear that our baseline models are not able to successfully leverage multimodal data to make predictions. Specifically, baseline models with handcrafted features have the following limitations:

1) The inability to extract handcrafted features from several trials (e.g., due to an insignificant SCR peak) reduces the amount of training data available to train both modality-specific and multimodal models.
2) Simple and interpretable models may fail to learn cross-modal information from feature combinations, especially with a limited dataset such as is common in affective computing studies, making it difficult to leverage multimodal data.

We propose to address these limitations by learning useful multimodal representations using a multitask supervised contrastive learning approach. These embeddings provide two clear advantages over handcrafted features:

1) By not relying on handcrafted features and instead using the raw time series data from each trial to extract meaningful representations, we are able to both learn the temporal dynamics of the signals within a window as well as utilize more trials and thereby increase our training data.
2) The proposed approach of supervised contrastive learning allows us to learn embeddings that combine information across modalities, leveraging complimentary sources of data better than baseline models using feature combinations.

The following section describes our approach in more detail and demonstrates how it makes the most of all three modalities to achieve better prediction performance.

V. Classification Using Multitask Supervised Contrastive Learning

In order to extract useful data representations for classifying participants with and without psychological disorders, we took inspiration from the supervised contrastive learning (SupCon) framework originally proposed for image classification [27]. Using this framework, we trained an *embedding model* that takes data from each trial as input and provides a corresponding multidimensional representation (or *embedding*) as the output. These embeddings can then be used as features for downstream tasks such as psychopathology prediction. Figure 2 provides an overview of our proposed approach.

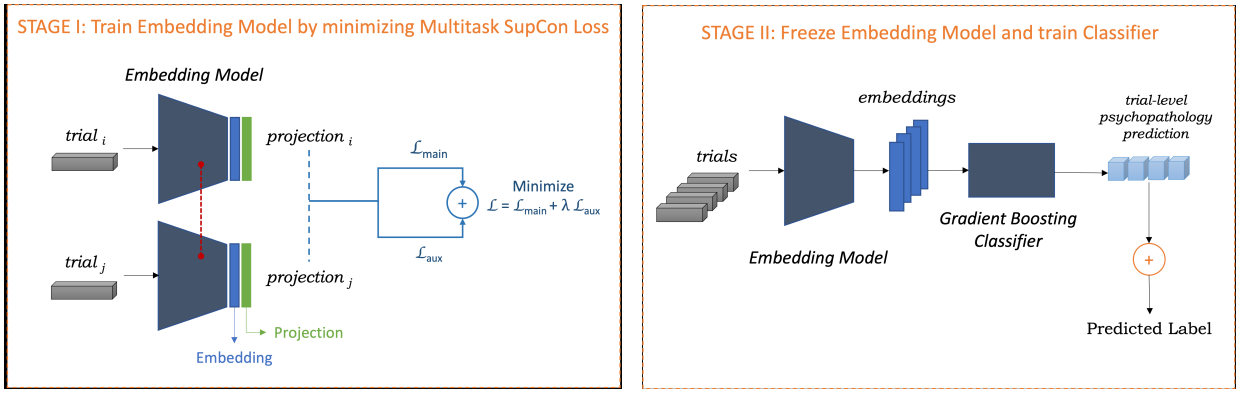


Fig. 2: An overview of the proposed approach. In the first stage, an embedding model is trained by minimizing the multitask supervised contrastive loss between the projections from a pairs of trials that have identical class labels. After training, the model is frozen and used to extract trial-level embeddings. A gradient boosting classifier is trained to classify the embeddings from each trial as clinical vs. non-clinical. Trial-level probabilities are averaged to obtain individual-level labels.

A. Supervised Contrastive Learning for Embedding Extraction

Recent approaches in self-supervised learning have used a contrastive loss to train models that minimize the distance between a given sample ("anchor") and a random augmentation of the sample ("positive") in the embedding space, while pushing away the "anchor" from other input samples (called "negatives"). Supervised contrastive learning extends this idea by utilizing the available labels of each sample, and minimizing the cosine distance between embeddings of two samples of the same class.

In this work, we employ this idea to first train an embedding model that minimizes the supervised contrastive loss based on psychopathology labels. This minimizes the distance between the embeddings of trials with identical labels (clinical or non-clinical). The embedding model contains an *encoder network* that maps the trial-level input data $\boldsymbol{x}$ to an embedding vector $\boldsymbol{r} \in \mathbb{R}^{\mathcal{D}_E}$, which is normalized to the unit hypersphere and will be used for downstream classification tasks. This is followed by a *projection network*, which maps the embedding $\boldsymbol{r}$ to a projection vector $\boldsymbol{z} \in \mathbb{R}^{\mathcal{D}_P}$. The projection vector $\boldsymbol{z}$ is also normalized to lie on the unit hypersphere and is used to calculate the distance between representations of different inputs. Mathematically, the embedding model (encoder + projection networks) is trained by minimizing the SupCon loss given by:

$$\mathcal{L} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)} \tag{1}$$

where i is an index within I, the set of all sample indices, and $A(i)$ is the set of indices from I excluding i. $P(i)$ is the set of all positive-labeled indices in the batch excluding i, and $|P(i)|$ is its cardinality. τ is the temperature hyperparameter which is set to 0.1 based on [27].

After the embedding model is trained, the projection network is discarded and the frozen encoder network is used to extract embeddings $\boldsymbol{r}_i$ for each trial i. These embeddings are used to train a gradient boosting classifier to predict psychopathology as described in Section IV-B.

B. Multitask SupCon with Auxiliary Labels

To better learn embeddings from trial-level data, we extended the supervised contrastive learning framework described above by using a multitask learning approach. In addition to learning from psychopathology labels, we added an auxiliary task to the embedding network by supervising the training using trial-level feedback labels corresponding to the positive or negative feedback received by participants. To this end, we first used the projection vector $\boldsymbol{z}$ to calculate the contrastive loss specified in Equation 1 using the psychopathology labels (clinical or non-clinical). This is henceforth referred to as $\mathcal{L}_{main}$. We used the same projection vector $\boldsymbol{z}$ to also calculate $\mathcal{L}_{aux}$, which is the contrastive loss supervised using the feedback labels (positive or negative feedback received from the customer during the trial). This auxiliary task provides additional supervisory signal to the model, allowing it to encode differences between positive and negative feedback instances in addition to psychopathology status. The objective function minimized by the embedding network is now

$$\mathcal{L} = \mathcal{L}_{main} + \lambda \mathcal{L}_{aux} \tag{2}$$

where λ is a tunable hyperparameter.

C. Embedding Architecture for Multimodal Inputs

We now discuss how supervised contrastive learning is extended to a multimodal setting by using a network architecture with modality-specific heads and late fusion for the embedding model. Figure 3 presents our proposed multimodal embedding network, which consists of three convolutional heads that process each input modality. The intuition behind having modality-specific heads instead of an early fusion approach was to account for the varying sampling rates of each signal.

The first of these heads takes as input the GSR signal, which consists of a single channel sampled at 1000 Hz. The second head processes the video data, which is represented by the

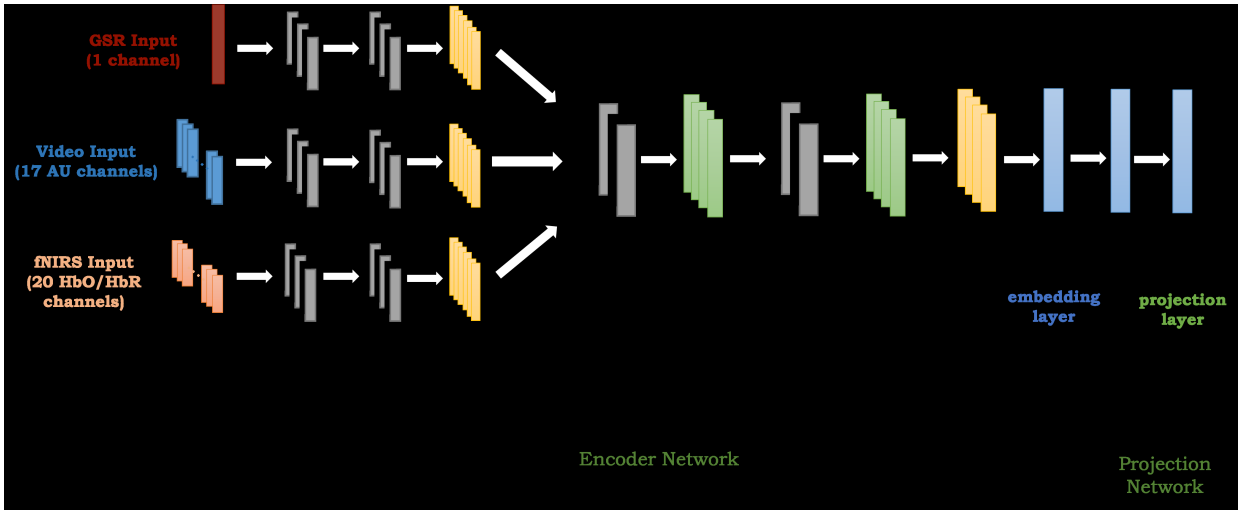


Fig. 3: Architecture of the multimodal embedding model with modality-specific heads and late fusion.

frame-by-frame intensity of the seventeen AUs extracted by OpenFace. We used the AUs as input rather than the raw video for two reasons – first, the presence of various AUs during emotion regulation has been linked to PFC neural activation and psychopathology in previous studies [14]. This led us to focus on AUs as specific biomarkers for psychopathology instead of allowing the network to extract other noisy or spurious features from facial video (e.g., learning facial structures or skin tones of participants and linking them to symptoms of dysregulation). Additionally, using AUs as input features preserves the privacy of participants by preventing the embedding network from encoding any features that could be used for facial recognition. Therefore, we used AUs sampled at 60 Hz as input to the second head in our multimodal architecture. The third input head processes the twenty fNIRS channels, corresponding to the ΔHbO and ΔHbR downsampled to 4 Hz from each of the ten source-detector pairs.

Each modality-specific head consists of a tunable number of 1D convolutional layers with ReLU activation and dropout, followed by an average pooling layer that downsamples the output of the last modality-specific layer to the same size for all inputs. These are then concatenated and passed through a shared convolutional network with 1D convolutions followed by batch normalization, ReLU, and dropout. The outputs at this stage are average pooled and flattened by passing through a fully connected layer and ReLU activation. This completes the encoder network architecture, and the outputs of this fully connected layer will be used as the embedding vector $\boldsymbol{r}$. During training, the model further contains a fully connected projection network with one hidden layer – the outputs $\boldsymbol{z}$ of the projection layer are used to compute the loss in Equation 2.

D. Training and Evaluation

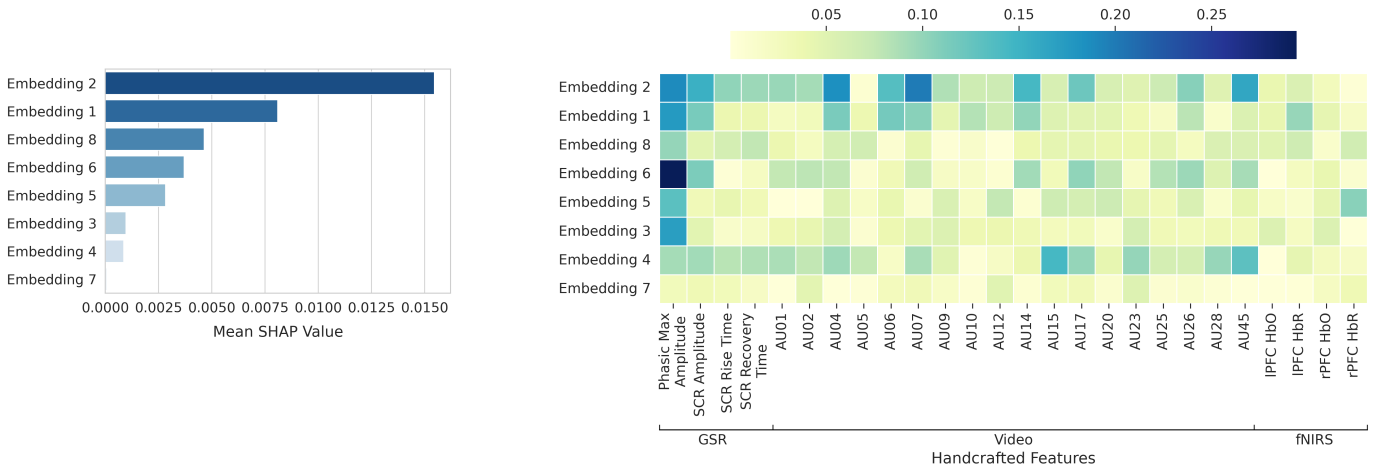
We now describe the process of training the embedding network (Stage I in Figure 2). We used the architecture described in Section V-C, with the number of convolutional layers, size of the convolutional and fully connected layers, convolutional kernel size, projection size, and dropout as tunable hyperparameters, along with the loss weight λ, learning rate, and batch size. We used the Adam optimizer to minimize the multitask supervised contrastive loss in Equation 2 and trained the network for up to 30 epochs, with early stopping if $\mathcal{L}_{main}$ failed to decrease by at least 0.001 in the last 10 epochs. The same training set and 3-fold cross validation strategy for hyperparameter tuning as described in Section IV-B were utilized for learning the embeddings. We thereby selected the model with lowest average $\mathcal{L}_{main}$ across all folds, which contains two convolutional layers with a kernel size of 2 and 16 output channels in the modality-specific heads, two convolutional layers with 8 output channels after combining modalities, followed by an embedding layer of size 8 and projection layer of size 16.

We then froze this embedding model and discarded the projection network, using the encoder network to extract multimodal embeddings for all trials. A gradient boosting classifier was trained in a manner similar to baseline models (Section IV-B) to predict psychopathology labels using the extracted embeddings as features.

TABLE II: Prediction performance using multimodal embeddings as features and comparison with baseline.

Modality	Embedding Length	Number of Valid Trials	Prediction Performance	
			AUROC	*F1 Score*
All	8	1413	0.90	0.91
Compare to Baseline		2.03x ↑	28% ↑	34% ↑

The prediction performance on the test participants is presented in Table II – our model is able to achieve an AUROC of 0.90 and an F1 score of 0.91. These results demonstrate a significant improvement over the baseline models (both modality-specific and multimodal), suggesting that embeddings learned through our multitask supervised contrastive learning approach capture informative features from multiple input signals.



(a) Mean SHAP values for each embedding feature for classifying psychopathology.

(b) Pearson correlation between embedding dimensions and handcrafted features from each modality.

Fig. 4: Interpreting predictions of the psychopathology classification model using multimodal embeddings in terms of mean SHAP values for embedding features and correlations between embeddings and handcrafted GSR, video, and fNIRS features.

VI. Discussion

Our proposed multitask supervised contrastive learning approach for classifying psychopathology addresses the important challenge of learning multimodal representations for classification tasks with limited data samples. We show that an embedding model with modality-specific heads and late fusion can learn representations that can be classified using a relatively lightweight and explainable model while achieving a high accuracy. This has important implications for the diagnosis of psychological disorders using behavioral and neurophysiological data from videos and wearable sensors.

While the multimodal embeddings generated through our multitask supervised contrastive learning approach are less interpretable than the handcrafted features extracted from each modality, they overcome two key challenges. First, we are able to train and extract embeddings for 1413 trials where raw data for all modalities were available, a 2.03x increase from the 696 training trials with handcrafted features across all modalities.

Posterior analysis of the embeddings also reveals more information about the information they capture – Figure 4 shows the importance of each embedding feature in terms of its mean SHAP value as well as the correlation between these features and the modality-specific handcrafted features described in Section IV-A. Note that the correlations shown in Figure 4b are across the 696 training trials where both handcrafted and embedding features are available. We observe that the most important embedding feature (embedding 2) is highly correlated with the presence of AU07, maximum phasic amplitude of the GSR signal, and AU04. We also note that different embedding features show high correlations with different handcrafted features across modalities, suggesting that the embedding model is able to learn cross-modal information. Our approach can therefore be applied to other scenarios with multi-sensor data to extract informative features.

While our model shows promising predictive performance when tested on the individuals in the held-out test set, our work is limited to a fairly small sample of individuals and should be tested in a broader population before it can be deployed for making diagnostic decisions. It is also imperative to test such models on different demographic groups to ensure fairness and reliability, though we attempt to minimize the possibility of encoding personal characteristics such as gender or race by using only desensitized AU features as inputs. The use of fNIRS as a modality may also limit the deployment of such systems for large-scale screening – future work could investigate ways to use a subset of modalities to make predictions when such models are deployed outside laboratory settings.

VII. Conclusion

This work presents a novel multitask supervised contrastive learning approach to extract multimodal embeddings of behavioral and neurophysiological data that can be used for classifying psychological disorder status in children as young as preschoolers. We show that the proposed approach identifies clinical participants with an AUROC of 0.90 using GSR, video, and fNIRS data collected during multiple trials of a clinically validated frustration-inducing task. The predictive performance improves significantly compared to models using baseline features from each modality, as well as their combination, demonstrating the utility of supervised contrastive learning in generating informative cross-modal features for downstream classification tasks. Our work also establishes the feasibility of identifying psychological disorders in early childhood using facial videos and wearable sensors while children complete a child-friendly emotion-inducing game, which has important implications for broadening access to mental health screening.

REFERENCES

[1] S. D. Calkins and S. Marcovitch, "Emotion regulation and executive functioning in early development: Integrated mechanisms of control supporting adaptive functioning.," 2010.

[2] P. Shaw, A. Stringaris, J. Nigg, and E. Leibenluft, "Emotion dysregulation in attention deficit hyperactivity disorder," *American Journal of Psychiatry*, vol. 171, no. 3, pp. 276–293, 2014.

[3] A. S. Grabell, Y. Li, J. W. Barker, L. S. Wakschlag, T. J. Huppert, and S. B. Perlman, "Evidence of non-linear associations between frustration-related prefrontal cortex activation and the normal: abnormal spectrum of irritability in young children," *Journal of abnormal child psychology*, vol. 46, no. 1, pp. 137–147, 2018.

[4] M. Kovacs, J. Sherrill, C. J. George, M. Pollock, R. V. Tumuluru, and V. Ho, "Contextual emotion-regulation therapy for childhood depression: Description and pilot testing of a new intervention," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 45, no. 8, pp. 892–903, 2006.

[5] P. A. Graziano, R. D. Reavis, S. P. Keane, and S. D. Calkins, "The role of emotion regulation in children's early academic success," *Journal of school psychology*, vol. 45, no. 1, pp. 3–19, 2007.

[6] N. Eisenberg and R. A. Fabes, "Emotion, regulation, and the development of social competence.," 1992.

[7] L. S. Wakschlag, P. H. Tolan, and B. L. Leventhal, "Research review:'ain't misbehavin': Towards a developmentally-specified nosology for preschool disruptive behavior," *Journal of Child Psychology and Psychiatry*, vol. 51, no. 1, pp. 3–22, 2010.

[8] J. J. Hudziak, W. Copeland, C. Stanger, and M. Wadsworth, "Screening for dsm-iv externalizing disorders with the child behavior checklist: a receiver-operating characteristic analysis," *Journal of child psychology and psychiatry*, vol. 45, no. 7, pp. 1299–1307, 2004.

[9] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.

[10] A. H. Bettis, T. A. Burke, J. Nesi, and R. T. Liu, "Digital technologies for emotion-regulation assessment and intervention: A conceptual review," *Clinical Psychological Science*, vol. 10, no. 1, pp. 3–26, 2022.

[11] A. Cubillo, R. Halari, A. Smith, E. Taylor, and K. Rubia, "A review of fronto-striatal and fronto-cortical brain abnormalities in children and adults with attention deficit hyperactivity disorder (adhd) and new evidence for dysfunction in adults with adhd during motivation and attention," *cortex*, vol. 48, no. 2, pp. 194–215, 2012.

[12] F. A. Cupaioli, F. A. Zucca, C. Caporale, K.-P. Lesch, L. Passamonti, and L. Zecca, "The neurobiology of human aggressive behavior: neuroimaging, genetic, and neurochemical aspects," *Progress in neuropsychopharmacology and biological psychiatry*, vol. 106, p. 110059, 2021.

[13] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal system.," 2017.

[14] A. S. Grabell, T. J. Huppert, F. A. Fishburn, Y. Li, H. M. Jones, A. E. Wilett, L. M. Bemis, and S. B. Perlman, "Using facial muscular movements to understand young children's emotion regulation and concurrent neural activation," *Developmental science*, vol. 21, no. 5, p. e12628, 2018.

[15] A. S. Grabell, T. J. Huppert, F. A. Fishburn, Y. Li, C. O. Hlutkowsky, H. M. Jones, L. S. Wakschlag, and S. B. Perlman, "Neural correlates of early deliberate emotion regulation: young children's responses to interpersonal scaffolding," *Developmental cognitive neuroscience*, vol. 40, p. 100708, 2019.

[16] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

[17] S. Mai, H. Hu, and S. Xing, "Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 481–492, 2019.

[18] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 2018, p. 2225, 2018.

[19] C. A. Torres, Á. A. Orozco, and M. A. Álvarez, "Feature selection for multimodal emotion recognition in the arousal-valence space," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 4330–4333, IEEE, 2013.

[20] H. Yu, T. Vaessen, I. Myin-Germeys, and A. Sano, "Modality fusion network and personalized attention in momentary stress detection in the wild," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, IEEE, 2021.

[21] D. Spathis, I. Perez-Pozuelo, S. Brage, N. J. Wareham, and C. Mascolo, "Learning generalizable physiological representations from large-scale wearable data," *arXiv preprint arXiv:2011.04601*, 2020.

[22] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomedical Signal Processing and Control*, vol. 70, p. 103029, 2021.

[23] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2018.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[25] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6329–6333, IEEE, 2021.

[26] X. Sun, J. Zeng, and S. Shan, "Emotion-aware contrastive learning for facial action unit detection," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 01–08, IEEE, 2021.

[27] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020.

[28] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal contrastive training for visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6995–7004, 2021.

[29] W. Song, J. Han, and H. Song, "Contrastive embeddind learning method for respiratory sound classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1275–1279, IEEE, 2021.

[30] B. Gunel, J. Du, A. Conneau, and V. Stoyanov, "Supervised contrastive learning for pre-trained language model fine-tuning," *arXiv preprint arXiv:2011.01403*, 2020.

[31] N. Henderson, W. Min, J. Rowe, and J. Lester, "Enhancing multimodal affect recognition with multi-task affective dynamics modeling," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, IEEE, 2021.

[32] Y. Yao, M. Papakostas, M. Burzo, M. Abouelenien, and R. Mihalcea, "Muser: Multimodal stress detection using emotion recognition as an auxiliary task," *arXiv preprint arXiv:2105.08146*, 2021.

[33] M. Zheng, F. Wang, S. You, C. Qian, C. Zhang, X. Wang, and C. Xu, "Weakly supervised contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[34] Y.-H. H. Tsai, T. Li, W. Liu, P. Liao, R. Salakhutdinov, and L.-P. Morency, "Learning weakly-supervised contrastive representations," *arXiv preprint arXiv:2202.06670*, 2022.

[35] M. Kalanadhabhatta, A. M. Santana, Z. Zhang, D. Ganesan, A. S. Grabell, and T. Rahman, "Earlyscreen: Multi-scale instance fusion for predicting neural activation and psychopathology in preschool children," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–39, 2022.

[36] G. J. DuPaul, R. Reid, A. D. Anastopoulos, M. C. Lambert, M. W. Watkins, and T. J. Power, "Parent and teacher ratings of attention-deficit/hyperactivity disorder symptoms: Factor structure and normative data.," *Psychological Assessment*, vol. 28, no. 2, p. 214, 2016.

[37] L. S. Wakschlag, S. W. Choi, A. S. Carter, H. Hullsiek, J. Burns, K. McCarthy, E. Leibenluft, and M. J. Briggs-Gowan, "Defining the developmental parameters of temper loss in early childhood: implications for developmental psychopathology," *Journal of Child Psychology and Psychiatry*, vol. 53, no. 11, pp. 1099–1108, 2012.

[38] T. M. Achenbach and L. A. Rescorla, *Manual for the ASEBA preschool forms and profiles*, vol. 30. Burlington, VT: University of Vermont, Research center for children, youth, & families, 2000.

[39] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.