Modeling Affective Distress from Video: A Modality-Aware Multi-Branch Learning Approach

Om Govind Jha
IISER Bhopal
Bhopal, India
om22@iiserb.ac.in

Nilesh Kumar Sahu
IISER Bhopal
Bhopal, India
nilesh21@iiserb.ac.in

Haroon R. Lone
IISER Bhopal
Bhopal, India
haroon@iiserb.ac.in

Abstract

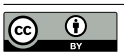
Automated detection of mental health conditions such as stress and Social Anxiety Disorder (SAD) from video is an important yet challenging problem, as these states exhibit both transient micro-expressions and sustained behavioral patterns over time. While stress and anxiety are distinct constructs, they share overlapping behavioral manifestations; in this work, we study them under a unified framework of affective distress. Existing video-based approaches often process heterogeneous behavioral cues using a single sequential encoder, which can lead to information loss when modeling signals with diverse temporal characteristics. To address this limitation, we propose a Multi-Branch Modality-Aware Multiple Instance Learning (MIL) framework that explicitly disentangles behavioral features based on their temporal dynamics. Specifically, high-variance and sparse signals (e.g., facial Action Units) are processed using a convolutional branch to capture transient behavioral patterns, while temporally continuous signals (e.g., postural landmarks and gaze) are modeled using a Bidirectional LSTM to capture long-term trends. A gated attention mechanism aggregates window-level representations into participant-level predictions while providing interpretable temporal saliency maps. We evaluate our approach across four datasets capturing stress- and anxiety-related conditions, including a private SAD dataset, StressID, UBFC, and AVEC 2017. Results demonstrate competitive or improved performance over state-of-the-art methods, along with strong cross-dataset generalization. Additionally, we introduce a gradient-based video rendering pipeline that overlays model attention onto input videos, enabling interpretable spatial and temporal insights for assessment. Our implementation is publicly available at *https://github.com/OmGovind15/MAMBL*.

CCS Concepts

• **Computing methodologies** → **Neural networks**; **Activity recognition and understanding**; • **Human-centered computing** → *Human computer interaction (HCI).*

Keywords

Affective Computing, Multiple Instance Learning, Video-based Behavior Analysis, Mental Health Assessment, Feature Routing



ICMI '26, Napoli, Italy

ACM ISBN 979-8-4007-2318-6/2026/10
https://doi.org/10.1145/3776574.3831196

ACM Reference Format:
Om Govind Jha, Nilesh Kumar Sahu, and Haroon R. Lone. 2026. Modeling Affective Distress from Video: A Modality-Aware Multi-Branch Learning Approach. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3776574.3831196

1 Introduction

Affective distress—including stress, anxiety, and emotion regulation difficulties—challenges clinical assessment and large-scale behavioral analysis. These conditions manifest through observable cues like facial expressions, gaze, and body movements, capturable non-invasively via video. Consequently, video-based analysis has emerged as a promising, scalable tool for mental health monitoring.

Recent advances in affective computing have demonstrated the potential of machine learning models to infer psychological states from multimodal behavioral signals [12, 13, 24]. In particular, video-based approaches leverage a rich behavioral vocabulary, including facial Action Units (AUs), head pose dynamics, eye gaze behavior, and upper-body kinematics. However, a key challenge lies in the heterogeneous temporal nature of these signals. Some behaviors, such as micro-expressions or rapid fidgeting, are transient and high-frequency, while others, such as posture or sustained gaze patterns, evolve gradually over longer durations. Treating these fundamentally different temporal dynamics within a single unified encoder can lead to suboptimal representations, where certain signal types dominate or obscure others.

A further practical challenge is that video recordings inherently vary in length across participants. A participant may be recorded for anywhere from a few seconds to over an hour, and diagnostically relevant events may occupy only a small fraction of the total duration. Standard sequence models that process fixed-length inputs are ill-suited to this setting. Multiple Instance Learning (MIL) provides a principled framework for handling such variability: each participant is treated as a *bag* of temporal window instances, and the model learns to assign discriminative importance to individual windows without requiring fine-grained annotations [7].

In this paper, we address these challenges with a framework we call **Multi-Branch Modality-Aware MIL**. Our key contributions are as follows: (i) **Principled Feature Routing:** We introduce a data-driven partitioning strategy that separates behavioral features into *sudden* (high-variance, rapidly changing) and *continuous* (temporally smooth and persistent) streams based on their temporal statistical properties. (ii) **Multi-Branch Encoding:** We design a dual-branch architecture consisting of a 1D Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (BiLSTM) network. The CNN is used to capture short-term,

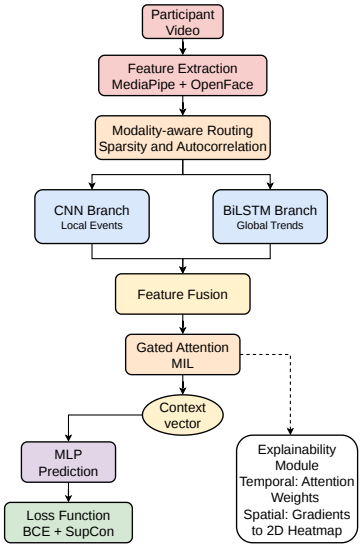


Figure 1: Overall framework of the proposed model architecture.

localized temporal patterns within small windows (e.g., rapid movements or micro-expressions), while the BiLSTM models longer-term temporal dependencies and gradual behavioral trends across the sequence. (iii) **Gated Attention MIL:** We employ a gated attention-based Multiple Instance Learning (MIL) mechanism [7] to aggregate window-level representations into a single participant-level prediction. This mechanism assigns interpretable importance weights to different temporal segments, enabling the model to focus on the most informative parts of the recording. (iv) **Explainability via Video:** We develop a gradient-based, attention-guided visualization pipeline that maps model importance scores back onto the original video frames. This produces continuous spatial and temporal overlays that help interpret which behaviors influence the model's predictions. (v) **Multi-Dataset Evaluation:** We benchmark our framework across multiple datasets capturing different aspects of affective distress, including our own private SAD dataset, StressID, UBFC, and AVEC 2017. Our results demonstrate strong performance and consistent generalization across datasets.

2 Related Work

Video-Based Affective Computing. Automated behavioral video analysis is central to affective computing. Early methods relied on hand-crafted descriptors like Local Binary Patterns and facial landmark statistics for depression and emotion recognition [3, 21]. Recent approaches leverage deep learning to extract joint spatio-temporal representations. RNNs and temporal CNNs frequently model dynamics from facial Action Units (AUs), head pose, gaze, and body movements [1, 4]. Furthermore, the AVEC challenges [21] have established crucial standardized benchmarks for multimodal affect analysis.

Temporal Modeling of Behavioral Signals. Modeling temporal dynamics is critical for understanding behavioral expressions of affective states. Prior work has employed recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, and spatial-temporal graphs to capture sequential dependencies in multimodal data [10, 11, 24]. These models are effective in learning temporal patterns but typically process all input features through a single encoding pathway. However, behavioral signals exhibit heterogeneous temporal characteristics. Some cues, such as micro-expressions or rapid movements, occur as short-lived, high-frequency events, while others, such as posture or gaze patterns, evolve gradually over longer time scales. Existing approaches that use a unified encoder may fail to adequately capture both types of dynamics simultaneously, leading to biased representations. Our work addresses this limitation by explicitly separating and modeling distinct temporal behaviors through a multi-branch architecture.

Multiple Instance Learning for Video-Based Analysis. MIL provides a natural framework for learning from weakly labeled sequential data, where labels are available only at the bag level [7]. In video-based settings, each recording can be treated as a bag of temporal segments, allowing models to identify salient moments without requiring frame-level annotations. Attention-based MIL methods have been particularly effective, as they enable interpretable aggregation of instance-level features. Ilse et al. [7] introduced a gated attention mechanism for MIL, which has since been widely adopted across domains. Despite these advances, most MIL-based approaches rely on a single feature representation pipeline and do not explicitly account for differences in temporal behavior across modalities. In contrast, our approach integrates modality-aware feature routing with MIL, allowing the model to better capture diverse behavioral patterns before aggregation.

Graph-Based and Transformer-Based Models. Graph-based methods model structured relationships between facial landmarks or body keypoints using Graph Neural Networks. Approaches such as ST-GCN [25] capture spatio-temporal dependencies in skeletal data and have been widely applied to action recognition and behavioral analysis. Extensions of these models have been explored for affective computing by modeling interactions between different facial and body regions. Transformer-based models have gained popularity for sequence modeling due to their ability to capture long-range dependencies through self-attention [22]. Time-series transformers [26] and multimodal transformer architectures [20] have shown strong performance in affective and stress prediction tasks. However, these models typically assume fixed-length inputs and process all features uniformly, without explicitly distinguishing between different temporal characteristics of behavioral signals. This can limit their effectiveness in real-world settings with variable-length recordings and heterogeneous features.

Explainability in Affective Computing. Interpretability is an important consideration in affective computing, particularly for applications related to mental health. Prior work has explored attention mechanisms and saliency-based techniques to identify which parts of an input signal contribute most to model predictions [7, 18]. In video-based settings, these approaches enable visualization of important temporal segments or spatial regions. While attention mechanisms provide coarse interpretability at the sequence level, gradient-based methods such as Grad-CAM [18] allow finer-grained localization of model activations. Our work builds on these ideas by

combining attention-based temporal weighting with gradient-based visualization, producing continuous spatial and temporal saliency overlays that enhance interpretability for behavioral analysis.

In summary, prior work has made substantial progress in video-based affective analysis using deep sequential models, MIL frameworks, and attention-based architectures. However, limited attention has been given to explicitly modeling the heterogeneous temporal characteristics of behavioral signals within a unified framework. Our proposed Multi-Branch Modality-Aware MIL approach addresses this gap by combining specialized feature encoding with interpretable instance-level aggregation, resulting in improved robustness and cross-dataset generalization.

3 Methodology

We extract a rich set of behavioral features from the participants' recordings and partition the timeline into N_i overlapping temporal windows of fixed duration L. Each participant, P_i is thus represented as a variable-length *bag* of instances: $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}\}$, where $x_{i,k} \in \mathbb{R}^d$ is the feature vector for the k-th temporal window of participant P_i. The participant label $y_i \in \{0, 1\}$ (e.g., non-stressed / stressed) is available only at the bag level; no per-window labels exist. The goal is to learn a mapping $f : \mathcal{X} \to \{0, 1\}$ that correctly predicts y_i from X_i, while simultaneously producing window-level attention weights $\alpha_{i,k}$ to serve as an interpretable importance map. Figure 1 provides an overview of the steps in our methodology.

3.1 Feature Extraction

To represent the video frames, we extract raw spatial landmarks and action units using the MediaPipe and OpenFace libraries. Because human behavior involves both static posture and dynamic motion, we do not rely solely on the raw frame-level values. Instead, we engineer a consistent set of statistical, kinematic, and spectral representations for every raw feature across the temporal window k. Let the continuous temporal sequence of a single raw feature within a window of length L be defined as $\mathbf{c} = [c_1, c_2, \ldots, c_L]$. To construct our final window-level vector $x_{i,k}$, we compute the following metrics: (i) **Base Statistics:** The mean ($\mu = \frac{1}{L}\sum_{t=1}^{L} c_t$) and standard deviation (σ) are calculated to capture the baseline activation and overall variance of the signal within the window, a standard practice in continuous affect modeling [10, 23]. (ii) **Kinematics (Derivatives):** To capture rapid micro-movements, psychomotor agitation (fidgeting), or tremors commonly associated with acute affective distress [17, 23], we calculate the temporal velocity $\mathbf{v}$ and acceleration $\mathbf{a}$ via finite differences, where $v_t = c_{t+1} - c_t$ and $a_t = v_{t+1} - v_t$. We extract the mean and standard deviation of both the velocity and acceleration. (iii) **Distributional Shape:** We compute the skewness (γ_1) and kurtosis (γ_2) to identify asymmetrical postural deviations and structural rigidity (postural freezing) [9, 17]: $\gamma_1 = \frac{\frac{1}{L}\sum_{t=1}^{L}(c_t-\mu)^3}{\sigma^3}$, $\gamma_2 = \frac{\frac{1}{L}\sum_{t=1}^{L}(c_t-\mu)^4}{\sigma^4}$. (iv) **Frequency Domain:** To capture cyclic physiological rhythms and repetitive behavioral motifs, such as respiration rate or involuntary postural swaying [6], we compute the Power Spectral Density (PSD) of the mean-centered sequence using Welch's method. Let f denote the frequency bins and $P(f)$ the corresponding power. From the PSD, we extract three distinct spectral descriptors: the dominant frequency ($f_{\text{dom}} = \arg\max_f P(f)$), the total spectral power ($P_{\text{total}} = \sum P(f)$), and the spectral centroid ($C = \frac{\sum f \cdot P(f)}{P_{\text{total}}}$).

3.2 Bimodal Feature Routing

Our architecture is built on the principle that distinct behavioral cues require specialized processing. Rather than treating all features equally, we use their temporal stability and frequency to group them into two dedicated processing streams. We define two metrics for every feature j: (i) **Sparsity (s_j):** The fraction of sequences in which feature j is inactive. (ii) **Autocorrelation (ρ_j):** The lag-1 temporal autocorrelation of feature j, measuring structural smoothness.

To be classified as a *Continuous* baseline shift (e.g., slouched posture or sustained gaze aversion), a feature must be both highly stable and frequently active. We enforce a logical intersection: For e.g., features satisfying both $\rho_j \geq 0.80$ and $s_j \leq 0.50$ are routed to **Stream B (Continuous)**. All other features—whether they are highly intermittent (high sparsity) or constantly erratic (low autocorrelation)—fail this strict criteria. These are classified as transient behavioral events (e.g., facial twitches, rapid fidgeting) and are explicitly routed to **Stream A (event)**. This intersectional routing strategy prevents modality collapse by ensuring that high-variance sudden movements do not mathematically overshadow subtle, continuous drifts during network training. The thresholds ($\rho = 0.80$, $s = 0.50$) were selected empirically based on preliminary validation experiments and were then kept fixed across all datasets. We further performed a sensitivity analysis by varying both thresholds. Performance remained relatively stable around the selected operating point, whereas more extreme settings reduced both accuracy and Macro-F1. In particular, overly restrictive sparsity thresholds occasionally caused feature partition collapse, routing nearly all features into a single stream.

3.3 Dual-Kinematic Feature Encoding

Each stream is processed by a dedicated neural branch architecturally matched to its temporal nature. **Stream A — 1D CNN :** To capture localized behavioral patterns, Stream A sequences are processed by a 1-Dimensional Convolutional Neural Network. Let $\mathbf{X}^A \in \mathbb{R}^{N_i \times d_A}$ represent the input sequence, where N_i is the number of temporal windows and d_A is the feature dimension. Unlike standard spatial CNNs, our filters convolve across the sequence dimension as $\mathbf{H}^A = \text{ReLU}(\text{BN}(\text{Conv1D}(\mathbf{X}^A)))$. Here, Conv1D extracts temporal features, BN denotes Batch Normalization to stabilize activations, and ReLU introduces non-linearity. By sliding across adjacent temporal windows, the CNN becomes highly sensitive to abrupt local transitions (e.g., sudden onset of hand-wringing). To ensure temporal synchronization for downstream fusion, appropriate padding is applied such that the output matrix $\mathbf{H}^A \in \mathbb{R}^{N_i \times d_{cnn}}$ preserves the original sequence length.

Stream B — Bidirectional LSTM Branch: To track long-range dependencies, Stream B sequences, denoted as $\mathbf{X}^B \in \mathbb{R}^{N_i \times d_B}$, are processed by a Bidirectional LSTM via $\mathbf{H}^B = \text{BiLSTM}(\mathbf{X}^B)$. The BiLSTM$(\cdot)$ function processes the sequence in both forward and backward directions, yielding a sequence of hidden states

$\mathbf{H}^B \in \mathbb{R}^{N_i \times d_{lstm}}$. This integration captures both historical and future contextual shifts, such as the gradual deterioration or stiffening of posture over the course of a video.

Feature Integration. To form a unified, multi-perspective representation for each temporal window, the output embeddings from both branches are dynamically concatenated. Let $\mathbf{h}_k^A \in \mathbb{R}^{d_{cnn}}$ and $\mathbf{h}_k^B \in \mathbb{R}^{d_{lstm}}$ represent the k-th temporal row of the intermediate matrices $\mathbf{H}^A$ and $\mathbf{H}^B$, respectively. The combined window-level embedding is defined as $\mathbf{h}_k = [\mathbf{h}_k^A; \mathbf{h}_k^B] \in \mathbb{R}^{d_h}$ where $[;]$ denotes the concatenation operator & $d_h = d_{cnn} + d_{lstm}$ is the dimension of the fused representation. This combined sequence is subsequently passed to the attention aggregation module.

3.4 Gated Attention Aggregation

To pool the variable-length sequence of window embeddings $\{\mathbf{h}_1, \mathbf{h}_2, ., \mathbf{h}_{N_i}\}$ into a single participant-level context vector, we employ a Gated Attention mechanism. Unlike standard soft attention, gated attention introduces a learned multiplicative gate that aggressively suppresses irrelevant windows. The attention score for the k-th window is formulated as $a_k = \mathbf{w}^\top(\tanh(\mathbf{V}\,\mathbf{h}_k) \odot \sigma(\mathbf{U}\,\mathbf{h}_k))$ where $\mathbf{w} \in \mathbb{R}^{d_a}$ is a learnable weight vector, and $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{d_a \times d_h}$ are learnable projection matrices mapping the input to an internal attention dimension d_a. The $\odot$ operator denotes element-wise multiplication, and σ is the sigmoid activation function acting as the gating mechanism. These scalar scores are normalized across the entire sequence via a softmax function to produce valid probability weights: $\alpha_k = \frac{\exp(a_k)}{\sum_{j=1}^{N_i} \exp(a_j)}$. The final, participant-level representation is the attention-weighted sum of all temporal windows: $\mathbf{z}_{\text{ctx}} = \sum_{k=1}^{N_i} \alpha_k\,\mathbf{h}_k \in \mathbb{R}^{d_h}$.

3.5 Classification and Contrastive Optimization

The pooled context vector $\mathbf{z}_{\text{ctx}} \in \mathbb{R}^{d_h}$ is passed through a Multi-Layer Perceptron (MLP) with dropout to yield the final binary prediction. The predicted probability $\hat{y} \in [0, 1]$ is given by $\hat{y} = \sigma(\mathbf{W}_2\,\text{ReLU}(\mathbf{W}_1\,\mathbf{z}_{\text{ctx}} + \mathbf{b}_1) + \mathbf{b}_2)$ where $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{b}_1, \mathbf{b}_2$ are learnable parameters, $\text{ReLU}(\cdot)$ introduces non-linearity, and $\sigma(\cdot)$ is the sigmoid activation function.

To train the network, we employ a hybrid loss that combines Binary Cross-Entropy (BCE) with Supervised Contrastive Loss: $\mathcal{L} = \lambda_{\text{BCE}}\,\mathcal{L}_{\text{BCE}} + \lambda_{\text{SupCon}}\,\mathcal{L}_{\text{SupCon}}$. The Binary Cross-Entropy loss is defined as $\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$ where $y_i \in \{0, 1\}$ is the ground-truth label and $\hat{y}_i$ is the predicted probability. To improve representation robustness under high inter-subject variability, we incorporate Supervised Contrastive Loss [8]. Let $\mathbf{z}_i$ denote the normalized embedding corresponding to sample i in a batch of size N. The contrastive loss is defined as $\mathcal{L}_{\text{SupCon}} = \sum_{i=1}^{N} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}$ where: (i) $P(i)$ denotes the set of indices of samples sharing the same label as i (excluding i itself), (ii) $A(i)$ denotes all samples in the batch excluding i, and (iii) τ is a temperature scaling parameter. This objective encourages embeddings of samples from the same class to be close in the latent space while pushing apart those from different classes. The balancing coefficients λ_{BCE} and λ_{SupCon} are determined empirically to achieve an effective trade-off between classification performance and representation structure.

3.6 Explainability: Attention Mapping

Beyond binary prediction, our architecture yields actionable spatial and temporal justification. The temporal attention weights $\{\alpha_1, \ldots, \alpha_{N_i}\}$ explicitly isolate the exact temporal windows driving the model's diagnosis. To achieve spatial explainability, we compute the gradient of the unnormalized prediction logit, y_{logit}, with respect to the input features for each temporal window k: $\mathbf{g}_k = \left|\frac{\partial y_{\text{logit}}}{\partial \mathbf{x}_k}\right|$. We compute gradients with respect to the pre-sigmoid logit to prevent gradient saturation. Because each dimension of $\mathbf{x}_k$ maps directly to a physical MediaPipe or OpenFace keypoint, we project the magnitude of $\mathbf{g}_k$ back onto the 2D image coordinates of the source video. To produce a continuous and readable visual heatmap, we apply a dual-smoothing strategy. Temporally, the attention weights α_k are linearly interpolated across the frames within each window to prevent visual flickering. Spatially, the gradient magnitudes $\mathbf{g}_k$ are multiplied by this interpolated attention weight, dynamically modulating the intensity of the spatial heatmap based on the model's moment-to-moment focus. This allows to observe exactly which sudden facial or postural shifts triggered the network's prediction at any given time.

4 Experiments and Results

4.1 Datasets and Preprocessing

We evaluate our framework on four datasets covering different aspects of affective distress, including stress, anxiety-related behavior, and depression-oriented benchmarks. While several of these benchmarks provide rich multimodal data (e.g., audio, physiological signals), our evaluation strictly isolates the video modality. This ensures a consistent, deployable setup to rigorously validate our video-based behavioral framework. To standardize the evaluation, all video sequences are processed into fixed-duration temporal windows from which behavioral features are extracted.

Private SAD Dataset. This dataset [17] consists of 101 participants collected in a controlled setting under a structured social stress paradigm (public speaking and social evaluation tasks). Each participant is assigned a binary label (SAD / non-SAD) based on assessment. Video recordings are segmented into temporal windows, yielding sequences with variable lengths ranging from 4 to 26 windows per participant (mean: 17.8). From each window, we extract a total of 2,736 features, split into 1,236 sudden (high-variance) features and 1,500 continuous features. This dataset has moderate class balance (44 SAD vs 57 non-SAD participants). Unlike other dataset feature processing, random forest pruning was not applied to the Private SAD dataset because its substantially larger sample size, together with the intrinsic regularization of the proposed network, sufficiently mitigated overfitting.

UBFC-Phys. UBFC-Phys is a multimodal dataset [15] designed for stress analysis. The elicitation protocol involves socio-evaluative stressors, typically public speaking or arithmetic tasks. To ensure high feature quality, the dataset (including all other datasets) underwent strict sparsity and variance thresholding prior to modeling. Behavioral features were extracted strictly from the video stream,

Table 1: Comprehensive Architectural Benchmark across Four Datasets (5-Seed Average). The best results per dataset are highlighted in bold.

Architecture	Private SAD		StressID		UBFC		AVEC 2017	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Modality-Aware Architectures								
Multi-Branch CNN-LSTM (Ours)	**71.04%**	**0.696**	70.74%	0.6968	71.67%	**0.635**	67.07%	0.562
FacialPulse MIL [23]	68.52%	0.678	70.71%	0.6964	71.33%	0.576	60.51%	0.559
FacialPulse Original [23]	60.90%	0.599	70.64%	0.6940	55.83%	0.445	62.42%	0.563
TMMT [10]	66.69%	0.647	70.43%	0.6927	70.67%	0.616	67.60%	0.513
SPG-GAT [24]	60.59%	0.548	**71.49%**	0.7007	59.17%	0.514	52.48%	0.406
Single-Stream Sequence Architectures								
DeepConvLSTM [11]	67.94%	0.663	71.44%	0.7034	70.83%	0.569	68.56%	0.542
TS-Transformer [26]	67.93%	0.661	70.43%	0.6927	67.67%	0.550	65.58%	**0.572**
CNN / ViT Baselines								
VideoMAE (ViT)[19]	54.32%	0.499	67.16%	0.640	37.30%	0.28	N/A‡	N/A‡
ResNet18-3D [5]	58.17%	0.366	65.28%	0.395	**77.78%**	0.63	N/A‡	N/A‡
ML Baselines (Temporal Pooling)								
XGBoost	59.48%	0.565	68.21%	0.673	43.33%	0.3860	61.01%	0.4940
Logistic Regression	61.43%	0.573	62.18%	0.611	45.00%	0.3567	53.15%	0.4216
SVM (RBF Kernel)	53.38%	0.386	69.18%	0.683	41.67%	0.3357	**69.68%**	0.4090
KNN	56.33%	0.509	65.82%	0.648	39.17%	0.3193	54.60%	0.4179
Random Forest	58.90%	0.533	71.16%	0.702	34.33%	0.2589	67.24%	0.4072
Ablation Variants (Investigating Routing Strategies)								
Baseline Dual-Stream (Linear)	65.14%	0.639	70.05%	0.6854	61.00%	0.495	64.33%	0.543
Unified 1D-CNN	65.53%	0.636	71.07%	0.7014	65.33%	0.521	62.93%	0.553
Unified BiLSTM	64.93%	0.636	71.39%	**0.7035**	66.33%	0.588	63.57%	0.517
No-Split Dual-Stream	65.53%	0.644	69.98%	0.6902	64.00%	0.521	68.48%	0.558
Learnable Routing	63.75%	0.623	69.88%	0.6834	69.83%	0.564	66.70%	0.547

‡ For AVEC 2017, raw video data was not available in our processed pipeline.

capturing facial and upper-body dynamics. The dataset contains 32 video sequences from 16 participants. Because UBFC-Phys contains only 16 participants, the expanded feature space was pruned using Random Forest feature importance to retain the 400 most informative features, thereby reducing overfitting. Each sequence consists of 35 temporal windows (uniform length), with each window represented by 400 features (375 sudden and 25 continuous). Based on the raw continuous anxiety scores, the class distribution is relatively balanced (56.25% high anxiety vs 43.75% low anxiety at the window level).

StressID. StressID is a publicly available multimodal dataset [2] containing synchronized video, audio, and physiological signals for stress detection. To elicit distress, participants underwent a series of cognitive load tasks and emotional stimulus viewings. We isolate the `binary-stress` labels, explicitly filtering out sequences with missing or unreleased ground-truth values to ensure evaluation integrity. The curated dataset contains 578 video sequences from 53 participants, with each sequence consisting of 72 temporal windows (uniform length). A similar Random Forest-based feature selection strategy was employed for StressID due to the limited number of participants. After preprocessing and feature selection, each window is represented by 400 features (149 sudden and 251 continuous). The dataset is moderately balanced at the window level (53.98% stress vs 46.02% non-stress).

AVEC 2017. AVEC 2017 [14] is a widely used benchmark focusing on depression state estimation from multimodal recordings. The videos feature participants engaging in human-agent interviews conducted by an animated virtual interviewer. Ground-truth formulation is structured as binary classification utilizing the standardized PHQ8 binary labels extracted across the training and development splits; participants belonging to hidden test sets with unreleased labels were systematically excluded and deduplicated. The final clean dataset consists of 139 participants, each represented by a fixed-length sequence of 17 temporal windows, resulting in 2,363 total windows. The dataset exhibits class imbalance, with 97 control and 42 depressed participants (corresponding to 69.78% and 30.22% at the window level), handled via dynamic BCE class weighting. Following our modality-aware design, features were partitioned based on their temporal characteristics, resulting in 30 sudden features processed by the CNN branch and 13 continuous features processed by the LSTM branch (43 features per window). For AVEC 2017, we used the pre-extracted features provided by the original benchmark without additional feature selection step using Random Forest.

Because affective distress manifests as both sudden expressions and sustained behavioral shifts, we standardize the continuous video streams into fixed-duration overlapping windows prior to feature extraction. To ensure sufficient temporal context for the sequential backbones, we utilize a 50% overlap (stride) across all datasets. For the Private SAD, UBFC-Phys, and AVEC 2017 datasets, we extract windows of 500 frames, yielding a temporal span of approximately 14 to 18 seconds per window depending on the camera framerate (e.g., 35 FPS for UBFC, 27 FPS for SAD). Unlike the higher-resolution data (27–35 FPS), the StressID features were aggregated at a lower temporal resolution of 5 FPS. To capture details despite this sparse sampling, we utilized 5-second time-bins for windowing. If a sequence's final window lacks sufficient frames, it is padded with zeros and masked during the attention calculation to prevent artifacting. Rather than relying on frame-level alignment across datasets, the proposed framework operates on sliding-window statistical descriptors. Consequently, differences in camera frame rates are absorbed during window-level aggregation, producing temporally comparable behavioral representations for the downstream sequential encoders.

4.2 Baselines

We compare our architecture against state-of-the-art (SOTA), sequential, and vision-based baselines. **Affective Models:** We evaluate against **FacialPulse Original** [23], an implementation of the FMMM architecture. It utilizes a dual-stream Bidirectional GRU to independently process two distinct behavioral representations: absolute features (the raw, static spatial state) and relative features (the temporal first-derivative, representing behavioral velocity), relying on global temporal pooling prior to late-fusion averaging. We also test **FacialPulse MIL**, which bridges this dual-stream extraction with our the MIL pooling. Next, we implement **TMMT** [10], which fuses features early through a shared LSTM and applies a Transformer Encoder utilizing a learnable class token (`[CLS]`) for task-specific sequence aggregation. Finally, we compare against **Spectral-Temporal GAT (ST-GAT)** [24], adapted from the original Spatial Pose Graph Attention Network (SPG-GAT). Rather than relying on rigid spatial topology, deep temporal features are transformed into the frequency domain via Fast Fourier Transform (FFT) to construct a fully connected spectral graph, processed by the original authors' GAT engine.

Deep Sequential Models: We include **DeepConvLSTM** [11], a standard sequential model utilizing four consecutive 1D convolutional layers followed by a two-layer BiLSTM. We also evaluate a **Time-Series Transformer** [26], which utilizes linear input projection & positional encoding prior to multi-head self-attention. Both models were purposefully adapted with our MIL attention head in place of standard flattening to ensure a fair, backbone-only comparison.

Vision and Classical Baselines: To demonstrate the limitations of purely spatial-temporal learning without explicit modality routing, we evaluate standard vision backbones operating directly on raw video sequences, specifically **VideoMAE** [19] and **ResNet18-3D** [5]. Lastly, we compare against classical machine learning baselines (**Random Forest, Logistic Regression, SVM, and XGBoost**) operating on globally pooled window-level feature vectors to demonstrate the need of sequence-aware modeling.

4.3 Experimental Setup

All deep learning models are trained using 5-fold GroupKFold cross-validation grouped by participant ID to prevent subject leakage. Within each fold, approximately 80% of the participants are used for training and the remaining 20% for validation. Early stopping is performed using the validation accuracy with a patience of 10 epochs, and the checkpoint with the best validation performance is retained for testing on the held-out fold. Nested cross-validation was not employed because of the limited number of unique participants in several datasets. Final performance is reported as the mean over five random seeds to account for initialization variance. All models are implemented in PyTorch and trained on a single NVIDIA GPU using Adam (10^{-4} learning rate, 10^{-4} weight decay).

4.4 Main Results

Across the diverse evaluation settings, our modality-aware architecture consistently demonstrates (see Table 1) strong performance, achieving state-of-the-art results on both the private SAD and UBFC datasets. On the primary SAD dataset, our model achieves the highest accuracy (**71.04%**) and Macro F1 (**0.696**). The substantial performance gap over the original single-stream FacialPulse baseline (60.90% accuracy) highlights the severe limitations of processing heterogeneous behavioral cues through a unified pathway. By explicitly isolating high-frequency transient signals from slow-moving postural shifts, our multi-branch approach successfully prevents the modality collapse and information decay observed in pure recurrent networks. This dominance extends to the UBFC dataset, where our architecture achieves a leading accuracy of **71.67%** and an F1 score of **0.635**, comfortably outperforming competitive multimodal models like TMMT.

Furthermore, our framework demonstrates robust cross-dataset generalization on StressID and AVEC 2017. On StressID, while spatial-temporal graph models (SPG-GAT) and deep recurrent networks (DeepConvLSTM) perform exceptionally well, our Multi-Branch architecture remains highly competitive (**70.74%** accuracy, **0.6968** F1), proving that it adapts effectively even when temporal characteristics vary. Similarly, on the AVEC 2017 benchmark—which introduces fixed-length sequences and significant class imbalance. Our model achieves an F1 score of **0.562**, closely trailing the TS-Transformer baseline.

Notably, across all datasets, classical machine learning baselines relying on global temporal pooling (e.g., XGBoost, Random Forest, SVM) exhibit high information loss. For instance, SVM collapses to an F1 score of 0.386 on SAD, and XGBoost drops to 0.3860 on UBFC. This widespread failure confirms that collapsing temporal windows washes out the critical transient micro-expression signals required for accurate affective distress detection, validating our choice of a MIL framework.

4.5 Architectural Ablations & Routing Analysis

To isolate the contribution of each architectural component, we evaluate simplified variants across all four datasets (Table 1). Overall, the results demonstrate that the proposed architecture benefits from both modality-aware branch specialization and statistically guided feature routing.

(**i**) **Effect of Branch Specialization:** Single-stream alternatives (Unified 1D-CNN and Unified BiLSTM) generally underperformed the proposed model, indicating that local temporal patterns and long-range dependencies are complementary and require dedicated CNN and recurrent processing.

(**ii**) **Importance of Statistical Feature Routing:** Dual streams alone are insufficient. While the **Baseline Dual-Stream (Linear)** preserves the statistical feature split but replaces the CNN and BiLSTM with simple linear encoders, the **No-Split Dual-Stream** retains the specialized branches but feeds the complete feature set to both. Both variants generally underperformed the proposed model, demonstrating that effective performance requires both statistically guided feature routing and branch specialization.

(**iii**) **Hard vs. Learnable Routing:** Replacing deterministic routing with Learnable Routing also reduced performance. The soft gating mechanism mixed heterogeneous feature types, limiting branch specialization. In contrast, routing using pre-computed statistical descriptors (Sparsity and Autocorrelation) preserved feature separation and produced more robust performance across datasets.

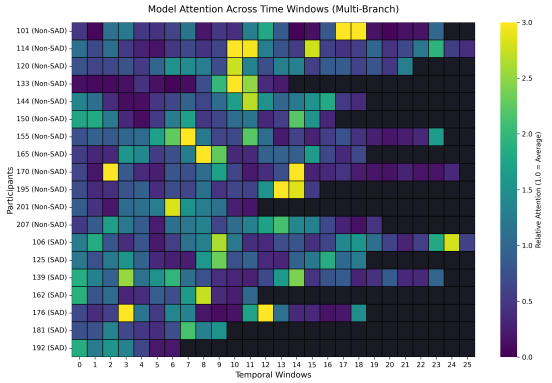


Figure 2: Temporal attention distribution across participants. Brighter values indicate higher attention weights $\alpha_{i,k}$.

4.6 Interpretability and Explainability

A key requirement in affective computing systems is transparency, i.e., understanding how and why a model arrives at a particular prediction. Interpretability refers to intrinsic transparency, where components of the model directly reveal how decisions are made. Explainability refers to post-hoc methods that attribute predictions to input features or regions. Our framework incorporates both: attention weights provide intrinsic interpretability, while gradient-based saliency enables post-hoc explainability.

Temporal Attention in MIL:. We had each participant P_i represented as a bag of temporal windows: $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,N_i}\}, \quad x_{i,k} \in \mathbb{R}^d$. Our model learns attention weights $\alpha_{i,k}$ over windows using a gated attention mechanism. Given window embeddings $h_{i,k}$, attention logits are computed as $a_{i,k} = \mathbf{w}^\top \left(\tanh(\mathbf{V}h_{i,k}) \odot \sigma(\mathbf{U}h_{i,k})\right)$. These logits are normalized across valid windows using a masked softmax: $\alpha_{i,k} = \frac{\exp(a_{i,k})}{\sum_{j=1}^{N_i} \exp(a_{i,j})}$. The participant-level representation is then obtained as a weighted aggregation: $z_i = \sum_{k=1}^{N_i} \alpha_{i,k}\, h_{i,k}$. This formulation ensures that $\alpha_{i,k}$ directly encodes the relative importance of each temporal window in the final prediction.

Temporal Attention Heatmaps: We visualize $\alpha_{i,k}$ across participants as a heatmap (Figure 2). Each row corresponds to a participant and each column to a temporal window. Brighter values indicate higher attention (i.e., windows contributing most strongly to the prediction), while darker regions correspond to suppressed windows. The visualization shows that attention is not uniformly distributed across time. Instead, the model assigns higher weights to a subset of temporal windows while down-weighting others. This selective allocation of attention reflects how the model prioritizes specific segments when forming its prediction.

From Temporal Attention to Frame-Level Saliency: While temporal attention provides coarse localization, it does not directly indicate which aspects of the input features are responsible for the prediction. To obtain finer-grained insight, we compute gradient-based saliency with respect to the input features. For a prediction $\hat{y}_i$, we compute $S_{i,k} = \left\| \frac{\partial \hat{y}_i}{\partial x_{i,k}} \right\|$. These gradients capture the sensitivity of the prediction to each temporal window. We combine them with attention weights to obtain a temporally weighted saliency signal: $\tilde{S}_{i,k} = \alpha_{i,k} \cdot S_{i,k}$. To produce frame-level saliency: (i) $\tilde{S}_{i,k}$ is interpolated from window-level resolution to frame-level timestamps. (ii) Feature gradients are mapped back to spatial coordinates using MediaPipe keypoints. (iii) A smoothing operation is applied to ensure temporal continuity.

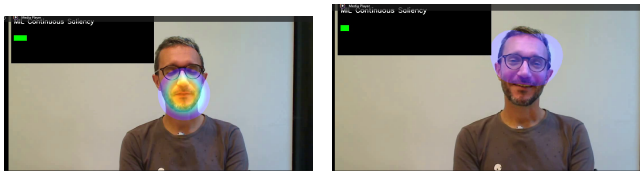


Figure 3: Video Saliency Visualization. Gradient-based saliency overlaid on video frames highlights spatial regions contributing to the model's prediction.

Implementation Details: Attention weights are obtained directly from the MIL forward pass, where masked softmax ensures that padded windows do not contribute. Gradients are computed via standard backpropagation, and saliency is accumulated per window before interpolation. The pipeline operates post-hoc and does not require additional supervision.

Video Saliency Visualization: The resulting saliency maps (Figure 3) overlay importance scores directly onto the video frames. This produces a continuous visualization of where the model is attending over time and which spatial regions contribute most to its prediction, enabling qualitative inspection of the model's behavior.

5 Discussion

Several recent works have explored multimodal and feature-driven approaches for anxiety and affect detection using video and sensor data. For instance, AnxietyFaceTrack [16] and related video-based studies [17] rely on handcrafted or learned facial and behavioral features combined within a unified modeling framework. Similarly, multimodal systems leveraging behavioral and physiological signals [9, 12] process different sources jointly, often through early or late fusion. Other works have focused on specific modalities such as gaze dynamics [13], demonstrating the predictive value of individual behavioral signals. While effective, these approaches typically treat features within a shared representation space without explicitly accounting for differences in their temporal behavior.

In contrast, our approach explicitly separates behavioral signals based on their temporal statistical characteristics prior to encoding. Empirically, this design leads to improved or comparable performance across multiple datasets. On the SAD and UBFC datasets, our model achieves the strongest results among all evaluated methods, outperforming both classical baselines and recent deep learning approaches such as FacialPulse [23] and attention-based multimodal models [10]. These gains are notable given that the competing models have comparable or greater architectural complexity.

Our primary contribution lies in rethinking how heterogeneous behavioral features are organized prior to learning. While prior work has explored attention-based Multiple Instance Learning [7], multimodal fusion [20], and sequence modeling independently, our framework combines (i) statistically grounded feature routing, (ii) modality-specific encoding, and (iii) attention-based aggregation within a unified pipeline. The empirical results suggest that this

combination provides a practical advantage in settings characterized by variable-length sequences and mixed temporal dynamics.

Relation to Explainability Approaches. Prior work on explainability in affective computing has largely relied on gradient-based visualization techniques such as Grad-CAM [18], which highlight spatial regions in individual frames that influence model predictions. While effective for spatial interpretation, these methods are typically applied post-hoc and do not explicitly capture temporal structure in sequential data. In contrast, our approach provides interpretability at both temporal and feature levels. The attention weights learned within the MIL framework directly indicate which temporal segments contribute most to the prediction, while gradient-based saliency is used to further localize feature-level importance within those segments. This combination enables a structured interpretation of model behavior over time, rather than relying solely on frame-wise explanations. Overall, our findings highlight that accounting for the temporal structure of behavioral signals can improve performance and robustness, while remaining competitive with state-of-the-art approaches across datasets.

Qualitative inspection of the generated saliency visualizations consistently revealed attention concentrated around the eye region, mouth, and mid-face. Moreover, these regions shifted dynamically across temporal windows as participant behaviour evolved. Although this qualitative analysis does not constitute clinical validation, the highlighted regions are broadly consistent with behavioural indicators commonly associated with affective distress, suggesting that the model attends to meaningful facial cues rather than background information.

Implications. (**i**) Our findings suggest that modeling heterogeneous temporal dynamics explicitly can be more effective than relying on a single unified encoder. Behavioral signals such as facial micro-movements and postural trends operate at different temporal scales, and treating them identically may lead to suboptimal representations. (**ii**) Our results indicate that architectural choices should be informed by dataset characteristics rather than applied uniformly. As observed across datasets, the relative benefit of feature routing depends on factors such as feature dimensionality, temporal resolution, and sequence variability. This suggests that simple diagnostic measures—such as feature sparsity and autocorrelation—can guide model selection in practice. (**iii**) the effectiveness of the MIL formulation reinforces the importance of handling variable-length sequences without requiring fine-grained annotations. By allowing the model to focus on informative temporal segments through attention, MIL provides a flexible and scalable alternative to fixed-length sequence modeling or global pooling approaches. (**iv**) the combined use of attention-based aggregation and gradient-based saliency highlights the value of multi-level transparency in sequence models. Providing both temporal and feature-level insights enables structured inspection of model behavior, which is important in applications involving behavioral analysis.

Why Feature Routing Matters? The performance trajectory across ablation variants tells a consistent story: the *source* of gains is not deeper architecture or larger capacity, but the principled disentanglement of behavioral signal types. Both the Unified CNN and Unified BiLSTM underperform the Multi-Branch model by 5–6% in accuracy, despite having comparable parameter counts. The No-Split Dual-Stream variant—which uses two branches but does not route features based on statistical properties—similarly underperforms, confirming that routing is the critical ingredient. Paradoxically, Learnable Routing performs worst: end-to-end soft routing allows the model to conflate the two regimes early in training, leading to degenerate solutions where both branches receive a mixed-type signal.

Dataset-Dependent Architecture Selection. Our cross-dataset analysis reveals an important practical insight: *the optimal architecture is not dataset-agnostic*. On high-dimensional, high-FPS data (SAD dataset), the Multi-Branch model with explicit routing is clearly superior. On lower-dimensional, lower-FPS data (StressID), pure recurrent models and even classical ML become competitive. This suggests that practitioners selecting architectures for affective computing tasks should consider the statistical profile of their feature space—specifically its dimensionality, temporal granularity, and feature-type composition—before defaulting to more complex models. Our statistical routing procedure is itself a diagnostic tool: if few features are classified as "sudden" (high sparsity), the benefits of the CNN branch are reduced, and a simpler unified model may suffice.

Limitations. The primary limitation of our approach is its reliance on controlled recording conditions (stable camera, face visibility, adequate lighting). In naturalistic settings, occlusions, variable head pose, and lighting changes may degrade MediaPipe landmark quality, propagating noise through all downstream features.

6 Safe and Responsible Innovation Statement

In alignment with ICMI's commitment to Safe and Responsible Multimodal Interaction, this framework is designed strictly as a decision support tool, mitigating the severe risk of autonomous misdiagnosis. The automated assessment of affective distress carries potential risks, including demographic bias, data privacy breaches, and misuse by non-medical personnel. To ensure responsible deployment and inclusivity, our architecture relies on continuous renderings to enforce a human-in-the-loop paradigm and transparently audit the evidence. All data was processed under strict privacy protocols. Real-world deployment mandates continuous auditing for bias mitigation to ensure equitable performance across diverse populations.

7 Conclusion

We proposed a Multi-Branch Modality-Aware Multiple Instance Learning framework for video-based affective distress detection. By routing features into sudden and continuous streams based on temporal statistics for specialized encoding (CNN and BiLSTM), our architecture prevents modality collapse. It achieves state-of-the-art performance on a private SAD dataset (71.04% accuracy, 0.696 Macro F1) and maintains robustness across diverse temporal granularities on different datasets. Furthermore, our gated attention and gradient-weighted overlay pipeline provides interpretable spatial-temporal saliency maps.

References

[1] Samuel Albanie, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. In *Proceedings of the 26th ACM International Conference on Multimedia* (Seoul, Republic of Korea) *(MM '18)*. Association for Computing Machinery, New York, NY, USA, 292–301. doi:10.1145/3240508.3240578

[2] Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma ISMAILOVA, Nicholas W., francois bremond, Massimiliano Todisco, Maria A Zuluaga, and Laura M. Ferrari. 2023. StressID: a Multimodal Dataset for Stress Identification. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 29798–29811. https://proceedings.neurips.cc/paper_files/paper/2023/file/5f09bfe6730e9627a9f800d01a8ad5cd-Paper-Datasets_and_Benchmarks.pdf

[3] Hamdi Dibeklioğlu, Zakia Hammal, Ying Yang, and Jeffrey F. Cohn. 2015. Multimodal Detection of Depression in Clinical Interviews. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (Seattle, Washington, USA) *(ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 307–310. doi:10.1145/2818346.2820776

[4] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016). https://api.semanticscholar.org/CorpusID:10075571

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), 6546–6555. https://api.semanticscholar.org/CorpusID:4539700

[6] Ruqiang Huang, Weihua Su, Shiyue Zhang, and Wei Qin. 2020. Remote Measurement of Vital Signs for Unmanned Search and Rescue Vehicles. In *2020 5th International Conference on Control and Robotics Engineering (ICCRE)*. 164–168. doi:10.1109/ICCRE49379.2020.9096468

[7] Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based Deep Multiple Instance Learning. (02 2018). doi:10.48550/arXiv.1802.04712

[8] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18661–18673. https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf

[9] Eunji Kim, Seungwan Jin, and Kyungsik Han. 2024. An Empirical Study on Social Anxiety in a Virtual Environment through Mediating Variables and Multiple Sensor Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 438 (Nov. 2024), 24 pages. doi:10.1145/3686977

[10] Giustino Marino, Alessandro Bria, Giuseppe Labanca, Paolo Soda, and Rosa Sicilia. 2025. Temporal Multimodal Multitask Attention for Affective State Estimation in a Stressful Environment. *IEEE Transactions on Affective Computing* (2025), 1–10. doi:10.1109/TAFFC.2025.3618385

[11] Francisco Javier Ordóñez and Daniel Roggen. 2016. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors* 16, 1 (2016). doi:10.3390/s16010115

[12] Jin-Hyun Park, Yu-Bin Shin, Dooyoung Jung, Ji-Won Hur, Seung Pil Pack, Heon-Jeong Lee, Hwamin Lee, and Chul-Hyun Cho. 2025. Machine learning prediction of anxiety symptoms in social anxiety disorder: utilizing multimodal data from virtual reality sessions. *Frontiers in Psychiatry* Volume 15 - 2024 (2025). doi:10.3389/fpsyt.2024.1504190

[13] Sabrina Patania, Alessandro D'Amelio, Vittorio Cuculo, Matteo Limoncini, Marco Ghezzi, Vincenzo Conversano, and Giuseppe Boccignone. 2023. Pain and Fear in the Eyes: Gaze Dynamics Predicts Social Anxiety from Fear Generalisation. In *Image Analysis and Processing - ICIAP 2023 Workshops - Udine, Italy, September 11-15, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 14365)*, Gian Luca Foresti, Andrea Fusiello, and Edwin R. Hancock (Eds.). Springer, 133–144. doi:10.1007/978-3-031-51023-6_12

[14] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (Mountain View, California, USA) *(AVEC '17)*. Association for Computing Machinery, New York, NY, USA, 3–9. doi:10.1145/3133944.3133953

[15] Rita Meziati Sabour, Yannick Benezeth, Pierre De Oliveira, Julien Chappé, and Fan Yang. 2023. UBFC-Phys: A Multimodal Database For Psychophysiological Studies of Social Stress. *IEEE Transactions on Affective Computing* 14, 1 (2023), 622–636. doi:10.1109/TAFFC.2021.3056960

[16] Nilesh Kumar Sahu, Snehil Gupta, and Haroon R. Lone. 2026. AnxietyFaceTrack: A Smartphone-Based Non-Intrusive Approach for Detecting State Anxiety Using Facial Features. *IEEE Transactions on Affective Computing* 17, 2 (2026), 2159–2170. doi:10.1109/TAFFC.2026.3666241

[17] Nilesh Kumar Sahu, Nandigramam Sai Harshit, Rishabh Uikey, and Haroon R Lone. 2024. Beyond questionnaires: Video analysis for social anxiety detection. *arXiv preprint arXiv:2501.05461* (2024).

[18] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. doi:10.1109/ICCV.2017.74

[19] Zhan Tong, Yibing Song, Jue Wang, Limin Wang, S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh. 2022-01-01. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training.

[20] Yao-Hung Tsai, Shaojie Bai, Paul Liang, J. Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting* 2019, 6558–6569. doi:10.18653/v1/P19-1656

[21] Michel Valstar et al. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of ACM Multimedia*.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[23] Ruiqi Wang, Jinyang Huang, Jie Zhang, Xin Liu, Xiang Zhang, Zhi Liu, Peng Zhao, Sigui Chen, and Xiao Sun. 2024. FacialPulse: An Efficient RNN-based Depression Detection via Temporal Facial Landmarks. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 311–320. doi:10.1145/3664647.3681546

[24] Jiaqi Xu, Hatice Gunes, Keerthy Kusumam, Michel Valstar, and Siyang Song. 2025. Two-Stage Temporal Modelling Framework for Video-Based Depression Recognition Using Graph Representation. *IEEE Transactions on Affective Computing* 16, 1 (2025), 161–178. doi:10.1109/TAFFC.2024.3415770

[25] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 912, 9 pages.

[26] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. 2021. A Transformer-based Framework for Multivariate Time Series Representation Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2114–2124. doi:10.1145/3447548.3467401