Multimodal Anxiety Detection with Adaptive Hyperbolic Few-Shot Learning

Aditya Sneh
IISER Bhopal
Bhopal, India
adityasneh2000@gmail.com

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

Anushka Sanjay Shelke
IISER Bhopal
Bhopal, India
anushka19@iiserb.ac.in

Arya Adyasha
IISER Bhopal
Bhopal, India
arya20@iiserb.ac.in

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

Abstract

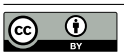
Anxiety disorders impact millions globally, yet traditional diagnosis relies on clinical interviews, while machine learning models struggle with limited labeled data. In this work, we introduce the Hyperbolic Curvature Few-Shot Learning Network (HCFSLN), a novel framework for multimodal anxiety detection using speech, physiological signals, and video. Our approach leverages hyperbolic embeddings to better capture hierarchical structure in affective signals, along with cross-modal attention and adaptive gating to model interactions across modalities. We collect a multimodal dataset from 108 participants and evaluate HCFSLN under few-shot settings. Our results show that HCFSLN significantly improves robustness and generalization, achieving 88% accuracy and outperforming baselines by 14%. Audio achieves the best unimodal performance (0.88 on social anxiety dataset, 0.79 on our dataset in 1-shot); adding physiological signals improves accuracy by ∼3–5%, and tri-modal fusion reaches up to 0.86 (social anxiety dataset) and 0.83 (our dataset), while using all modalities slightly reduces performance (∼0.79, ∼0.75) due to noise. The adaptive gating mechanism provides interpretability by weighting modalities, consistently prioritizing audio and leveraging physiological signals, with t-SNE showing clearer class separation. Further analysis highlights complementary multimodal cues for anxiety detection. Our approach is effective in low-data settings. We release our dataset and code at https://huggingface.co/datasets/SIR-Lab/Multi-Modal-Anxiety-Dataset.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Neural networks**; • **Human-centered computing** → *Empirical studies in HCI*.

Keywords

Multimodal Interaction; Affective Computing; Anxiety Detection; Mental Health; Few-Shot Learning; Representation Learning



ICMI '26, Napoli, Italy

ACM ISBN 979-8-4007-2318-6/2026/10
https://doi.org/10.1145/3776574.3831191

ACM Reference Format:
Aditya Sneh, Nilesh Kumar Sahu, Anushka Sanjay Shelke, Arya Adyasha, and Haroon R. Lone. 2026. Multimodal Anxiety Detection with Adaptive Hyperbolic Few-Shot Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3776574.3831191

1 Introduction

Anxiety disorder is one of the most prevalent mental health conditions, affecting over 301 million people worldwide [14]. Despite its widespread prevalence and substantial impact, detection still primarily relies on clinical diagnosis [38], which often requires multiple visits and can be influenced by variability in clinical judgment and diagnostic criteria.

Recent studies have investigated the use of wearable sensors for detecting and predicting mental health conditions by monitoring physiological signals [31, 32]. These sensors capture physiological data, which, when combined with self-reported measures, serve as inputs for machine learning and deep learning models. In these studies, participants engage in controlled anxious activities while physiological signals, such as heart rate variability and skin conductance, are recorded. Beyond physiological signals, researchers have also integrated multimodal data sources, including audio and video, to enhance mental health predictions [43]. However, many of these studies are constrained by small or imbalanced datasets, which increases the risk of overfitting [13]. Furthermore, publicly available multimodal datasets for mental health research remain scarce, hindering reproducibility and large-scale advancements in the field.

Existing studies on machine learning model performance suggest that training robust mental health prediction models requires at least 1,000 participants and 100,000 data points to ensure effective learning and evaluation [44]. However, collecting such large datasets is both time-consuming and financially burdensome. While data augmentation techniques are commonly used to compensate for small datasets, their application to physiological signals and multimodal data (such as audio and video) can introduce artifacts, distort underlying patterns, and ultimately degrade model performance [7]. To address the challenge of small training samples in anxiety classification, we explore Few-Shot Learning (FSL), a technique that enables models to achieve high accuracy with minimal data [8].

To demonstrate the usability of FSL and address the lack of publicly available datasets combining physiological and audiovisual data, we conducted a study in an Asian country to highlight FSL's relevance in anxiety research. We collected data from 108 participants, who performed a speech activity in front of a smartphone camera (see Figure 2). The dataset includes physiological signals such as Photoplethysmography (PPG) and Electrodermal Activity (EDA) from wearable sensors along with audiovisual recordings. Additionally, self-reported assessments were collected to label participants as anxious or non-anxious.

Moreover, this paper presents, Hyperbolic Curvature Few-Shot Learning Network (HCFSLN), a novel FSL framework with Hyperbolic Curvature for anxiety detection. HCFSLN enhances feature separability through hyperbolic embeddings, cross-modal attention, and an adaptive gating network, enabling robust classification with minimal data. It outperforms six state-of-the-art (SOTA) FSL baseline models on two different datasets, achieving 88% accuracy in the 1-shot setting with audio modality. These results highlight the advantage of hyperbolic space in modeling anxiety-related speech patterns. Our key contributions are:

- We propose **HCFSLN** and systematically benchmark recent few-shot learning models using Euclidean, hyperbolic, and hyperspherical embeddings, and demonstrate that our proposed hyperbolic-based framework consistently outperforms these baselines. To foster reproducibility, we release our code and dataset[1].
- We present the **Multi-Modal Anxiety Dataset (M2AD)**, a real-world dataset comprising physiological (PPG, EDA) and audiovisual signals from 108 participants in a low-middle-income country. This resource addresses a key gap in anxiety research and is made publicly available for further study.

2 Related Work

Studies on anxiety detection have explored various modalities, like EDA, Electrocardiogram (ECG), audio, video, etc., to capture behavioral and physiological cues [27]. Electroencephalogram (EEG) and ECG provide detailed neural and cardiac insights but require costly equipment and controlled conditions, limiting scalability. However, wearable sensors like PPG and EDA offer a low-cost, unobtrusive solution for continuous physiological monitoring. Moreover, modalities like audio and video from ubiquitous audiovisual cameras present behavioral features.

Various deep-learning algorithms have been explored for anxiety detection. For example, EEG-based approaches use models like EmotionNet [4], a CNN-LSTM hybrid to capture anxiety-related brain activity. ECG based methods analyze heart rate variability using techniques such as probabilistic binary pattern extraction [2] and tunable q-factor wavelet transforms [15]. Similarly, [26] proposed MMD-AS, which employs an Improved Fireworks Algorithm for feature selection in smartphone-based anxiety detection. Likewise, [36] proposed speech-based models like AudiBERT, which incorporate acoustic features for voice-based mental health detection. However, to train these deep learning models, we need a large annotated dataset.

To address the limitations of small datasets in the mental health domain, FSL has emerged as a promising approach in mental health assessment, enabling classification with minimal labeled data. Recent studies have employed FSL techniques to detect stress and anxiety conditions. For example, a Siamese Neural Network with Wasserstein distance minimization has been applied to stress classification [9], effectively handling distribution mismatch across individuals. Similarly, the Spatiotemporal Feature Fusion for Detecting Anxiety framework, which integrates 3D-CNN + LSTM for feature extraction, enhances multimodal anxiety detection by leveraging few-shot learning for improved generalization [27]. Additionally, a meta-learning-based stress category detection framework use an encoder module with a Dependency Graph Convolutional Network, an induction module using a Mixture of Experts' mechanisms, and a relation module for similarity assessment, demonstrating strong generalization in low data scenarios [40].

Although FSL research has explored various embedding spaces—including Euclidean [19], hyperbolic [29], and hyperspherical [6] there remains a significant gap in understanding which geometry works best for physiological-only, audiovisual-only, or multimodal anxiety data across different shot scenarios. A study show that hyperbolic embeddings do not always outperform Euclidean ones; performance often depends heavily on training configuration and embedding radius [29]. While surveys emphasize hyperbolic geometry's strength in representing hierarchical or low-dimensional structures and its robustness in low-sample settings [25], they offer limited guidance on modality-specific embedding choices. Consequently, no prior work clarifies which embedding space works best for physiological-only, audiovisual-only, or multimodal anxiety data, especially under different shot scenarios. Our study fills this gap via a comprehensive comparison across embeddings, modality sets, and shot configurations.

3 Proposed Method

This section discusses HCFSLN (see Figure 1), a hyperbolic curvature-aware framework for multimodal FSL, leveraging hyperbolic embeddings, cross-modal attention, and adaptive gating to improve feature fusion and class separability.

3.1 Preliminaries

FSL classifies new samples with minimal labeled data using **episodic training**, where a model learns from a sequence of binary (i.e., anxious or non-anxious) classification tasks. Each episode consists of a support set The support set is defined as $\mathcal{S} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{N_S}$ with $N_S = 2K$ labeled examples (i.e., K examples per class), and the query set is $\mathcal{Q} = \{(\mathbf{x}_k, y_k)\}_{k=1}^{N_Q}$ with $N_Q = 2B$ unlabeled examples (i.e., B examples per class). Here, $\mathbf{x}_j \in \mathbb{R}^d$ is a d-dimensional multimodal feature vector, and $y_j \in \{0, 1\}$ is its binary class label. Similarly, $\mathbf{x}_k$ is a d-dimensional query feature vector with its corresponding label y_k. During training, y_k are available for loss computation, but during inference, FSL models predict query labels without supervision. Each data sample $\mathbf{x}$ consists of features from M different modalities. It is represented as

$$\mathbf{x} = \left[\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M)}\right] \tag{1}$$

[1] https://huggingface.co/datasets/SIR-Lab/Multi-Modal-Anxiety-Dataset

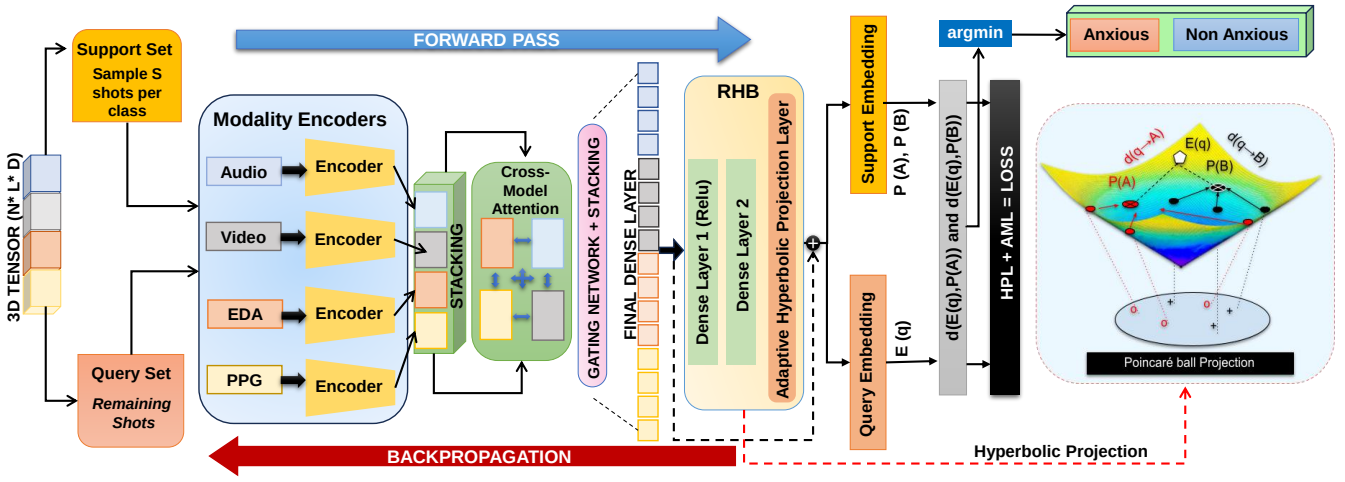


Figure 1: Architecture of the proposed HCFSLN framework.

where $\mathbf{x}^{(m)}$ is the feature vector from the m-th modality, such as audio, PPG, EDA, or video. The challenge lies in integrating heterogeneous modalities while preserving discriminative representations in small dataset settings and mitigating modality-specific noise and cross-modal interference. To address this, we learn an embedding function $f_\theta : \mathbb{R}^d \to \mathbb{B}^d$ that maps features from Euclidean space $\mathbb{R}^d$ to a hyperbolic space $\mathbb{B}^d$ (Poincaré ball model) to promote compact intra-class clustering and inter-class separation. A **modality cross-attention mechanism** with a **gating network** (filters noise) mitigates irrelevant noise during training [16]. Unlike Euclidean-based FSL models, our approach employs an **adaptive curvature parameter** (controls hyperbolic geometry) denoted by $\alpha > 0$ to capture complex non-Euclidean relationships. The objective is to develop a robust few-shot framework that integrates heterogeneous multimodal features in hyperbolic space, leveraging adaptive curvature and prototype-based classification for enhanced clustering and separation.

3.2 Multimodal Feature Fusion

Before any hyperbolic processing or prototype computation, the model obtains a unified multimodal representation $\mathbf{h}$ that effectively integrates all modalities.

3.2.1 Cross-Modal Attention. Each modality $\mathbf{x}^{(m)}$ undergoes feature extraction using dedicated **modality encoders**, which capture meaningful representations from raw multimodal inputs. Given an input sample with multiple modalities, the initial feature vector shown in Eq (1), where each $\mathbf{x}^{(m)}$ is processed through a modality-specific encoder $f_m(\cdot)$, which includes two convolutional layers (Conv1 with kernel size 3, Conv2 with kernel size 5) to capture temporal/spatial dependencies, followed by a dense layer with ReLU activation and dropout for dimensionality reduction and regularization. The dense output is expanded and passed to a MultiHeadAttention layer to re-weight features, and a residual connection adds this attention output back to the original features, followed by layer normalization to ensure stable training and effective information preservation [42]. The encoded modality representations are

$$\mathbf{h}^{(m)} = f_m(\mathbf{x}^{(m)}) \tag{2}$$

where $\mathbf{h}^{(m)}$ represents the high-level feature embedding of the m-th modality. To integrate complementary information, we first stack the modality-specific embeddings along a new axis, resulting in a tensor of shape (M, d'), where M is the number of modalities and d' is the embedding dimension. A cross-modal attention mechanism is then applied by computing dependencies between all modalities to refine this representation which enhances inter-modal interactions while suppressing noise from less informative modalities.

3.2.2 Gating Network. Following attention refinement, an adaptive gating network assigns importance scores to each modality to ensure that more informative modalities contribute more significantly to the final representation [45]. From Eq. (2), the fused multimodal representation is computed as

$$\mathbf{h} = \sum_{m=1}^{M} w_m \, \mathbf{h}^{(m)}, \tag{3}$$

where the learnable weight w_m for the m-th modality is defined as

$$w_m = \frac{\exp(\mathbf{W}\,\mathbf{h}^{(m)})}{\sum_{j=1}^{M} \exp(\mathbf{W}\,\mathbf{h}^{(a)})}$$

Here, $\mathbf{W}$ is a learnable parameter matrix that transforms each modality's embedding $\mathbf{h}^{(m)}$ into a scalar score, $\mathbf{h}^{(a)}$ represents the embeddings from all modalities, iterating over $a \in \{1, 2, \ldots, M\}$ and w_m is the normalized weight (via softmax) representing the relative importance of the m-th modality. This mechanism effectively tackles the challenge of **heterogeneous feature distributions**. The final representation $\mathbf{h}$ is then passed to the hyperbolic projection stage.

3.3 Adaptive Hyperbolic Projection

The unified multimodal representation $\mathbf{h}$ is now embedded into the **Poincaré ball** $\mathbb{B}^d$, a hyperbolic space that provides improved

geometric separation for classification [22]. Given the fused representation $\mathbf{h} \in \mathbb{R}^d$, the projection into hyperbolic space is performed as

$$\mathbf{y} = \tanh\big(\alpha \, \|\mathbf{h}\|\big) \cdot \frac{\mathbf{h}}{\|\mathbf{h}\|}$$

where $\alpha > 0$ is a trainable curvature parameter and y is the hyperbolic embedding of the multimodal representation $\mathbf{h}$. This mapping ensures that the output $\mathbf{y}$ lies within the Poincaré ball $\mathbb{B}^d$. Unlike fixed-curvature approaches, our model optimizes α dynamically during training, allowing it to adapt to the data distribution and curvature of the embedding space. Mapping the fused multimodal representation into a shared hyperbolic space ensures that intermodal relationships are effectively preserved [12].

3.3.1 Distance Computation in Hyperbolic Space. Once projected, embeddings are compared using the **Poincaré distance** [30], which measures similarity in hyperbolic space and replaces the standard Euclidean metric used in traditional prototypical networks. The distance between two embeddings $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{B}^d$ is computed as:

$$d_{\mathbb{B}}(\mathbf{y}_1, \mathbf{y}_2) = \operatorname{acosh}\left(1 + \frac{2\,\|\mathbf{y}_1 - \mathbf{y}_2\|^2}{\big(1 - \|\mathbf{y}_1\|^2\big)\big(1 - \|\mathbf{y}_2\|^2\big)}\right) \tag{4}$$

Here, $\operatorname{acosh}(\cdot)$ is the inverse hyperbolic cosine function. This formulation ensures that distances grow exponentially as points move toward the boundary of $\mathbb{B}^d$, enhancing class separability.

3.3.2 Residual Hyperbolic Block. A Residual Hyperbolic Block (RHB) refines the learned embedding by combining the original embedding with a transformed version of itself. Starting with an initial hyperbolic embedding $\mathbf{y}$, a trainable function $f(\mathbf{y})$ (e.g., a fully connected layer) is applied to capture high level features. The block then adds the original embedding back to this transformed output, resulting in

$$\mathbf{y}' = f(\mathbf{y}) + \mathbf{y}$$

Since we operate in the Poincaré ball $\mathbb{B}^d$ (a hyperbolic space), the output $\mathbf{y}'$ is typically re-projected into $\mathbb{B}^d$ to maintain stability. This residual enables the model to learn richer and more expressive embeddings while ensuring the geometric constraints of hyperbolic space are preserved.

3.4 Prototype Computation in Hyperbolic Space

Once both support and query samples have been embedded into the hyperbolic space, prototype-based classification is performed. For a given class c (with $c \in \{0, 1\}$), first compute the unweighted prototype:

$$\bar{\mathbf{p}}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{y}_i, \tag{5}$$

where $\mathbf{y}_i$ is the hyperbolic embedding of the i-th support sample and N_c is the number of support samples in class c. Then, the class prototype $\mathbf{p}_c$ is computed as a weighted mean in the hyperbolic space from Eq. (5) :

$$\mathbf{p}_c = \sum_i w_i \, \mathbf{y}_i, \quad w_i = \frac{\exp\Big(-d_{\mathbb{B}}(\mathbf{y}_i, \bar{\mathbf{p}}_c)\Big)}{\sum_j \exp\Big(-d_{\mathbb{B}}(\mathbf{y}_j, \bar{\mathbf{p}}_c)\Big)} \tag{6}$$

The weights w_i (obtained via a softmax over negative distances) ensure that points closer to the prototype contribute more. This formulation prevents prototype distortion due to outliers & maintains class consistency in hyperbolic space.

3.5 Loss Functions and Optimization

To enhance class separability in hyperbolic space, we employ a combination of hyperbolic prototypical loss and angular margin loss, optimizing decision boundaries for few-shot learning.

3.5.1 Hyperbolic Prototypical Loss. Given a query embedding $\mathbf{y}_q$ (representing a query sample in hyperbolic space) and class prototypes $\mathbf{p}_c$ (the representative embedding for class c), classification is performed by measuring the Poincaré distance $d_{\mathbb{B}}(\mathbf{y}_q, \mathbf{p}_c)$ (with $d_{\mathbb{B}}$ denoting the distance in the Poincaré ball) between the query and each prototype:

$$d_{\mathbb{B}}(\mathbf{y}_q, \mathbf{p}_c) = \operatorname{acosh}\left(1 + \frac{2\,\|\mathbf{y}_q - \mathbf{p}_c\|^2}{\big(1 - \|\mathbf{y}_q\|^2\big)\big(1 - \|\mathbf{p}_c\|^2\big)}\right) \tag{7}$$

The probability of a query belonging to class c is then obtained via a softmax over negative distances.

$$p(y_q = c \mid \mathbf{y}_q) = \frac{\exp\big(-d_{\mathbb{B}}(\mathbf{y}_q, \mathbf{p}_c)\big)}{\sum_{c' \in \{0,1\}} \exp\big(-d_{\mathbb{B}}(\mathbf{y}_q, \mathbf{p}_{c'})\big)} \tag{8}$$

The prototypical loss is defined as:

$$\mathcal{L}_{\text{proto}} = -\mathbb{E}\left[\log p(y_q = c \mid \mathbf{y}_q)\right], \tag{9}$$

which ensures that query embeddings remain close to their corresponding class prototypes while maximizing inter-class distances [11].

3.5.2 Angular Margin Loss. To further enhance class discrimination, an angular margin loss enforces a margin γ between the cosine similarity of a query with its correct prototype and the most similar incorrect prototype:

$$\mathcal{L}_{\text{angular}} = \mathbb{E}\left[\max\left(0, \cos(\theta_{q,p_c}) + \gamma - \max_{c' \neq c} \cos(\theta_{q,p_{c'}})\right)\right] \tag{10}$$

Here, c' indexes all classes other than the correct class c, and θ_{q,p_c} represents the angular distance between the query embedding $\mathbf{y}_q$ and the correct prototype p_c. The term $\max_{c' \neq c} \cos(\theta_{q,p_{c'}})$ selects the most similar incorrect prototype, enforcing an angular margin γ between classes. We adopt this loss from ArcFace [5] by incorporating an additive margin γ to increase inter-class separation while preserving intra-class compactness. The total loss is a weighted combination of the hyperbolic prototypical loss and the angular margin loss from Eqs. (9) and (10):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{proto}} + \lambda \, \mathcal{L}_{\text{angular}}, \tag{11}$$

where λ is a weighting parameter that controls the angular margin loss. This combined loss encourages structured hyperbolic embeddings with better class separability.

4 Experiments

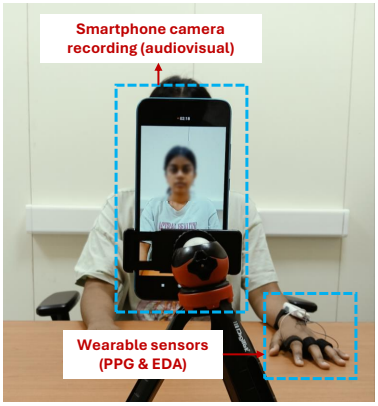


Figure 2: Experimental setup.

4.1 Datasets

Multi-Modal Anxiety Dataset (M2AD): The M2AD dataset was collected at IISER Bhopal, with ethical approval obtained from the institutional review board. In the study, student participants were invited to a laboratory setting to perform an anxiety-inducing speech task in front of a smartphone camera. Public speaking was chosen as the stressor based on the Trier Social Stress Test (TSST), a widely used protocol for eliciting psychological stress [17]. Participants were informed that they would speak for two minutes and that their recordings would be evaluated by experts.

On the day of the study, each participant wore Shimmer sensors on their fingers to collect PPG & EDA signals. Simultaneously, a smartphone (Redmi 9A), mounted on a tripod, recorded video of the participant (see Figure 2). A research assistant randomly selected a topic from a psychiatrist-approved list and instructed the participant to speak immediately, without any preparation time. The speech duration was fixed at two minutes. Example topics included: (i) the importance of following rules in social life, (ii) whether public transport should be free, and (iii) whether it is fair to apply the same grading system to all students.

Following the task, participants completed the anxiety subscale of the Depression, Anxiety, and Stress Scale (DASS-21) [21], which includes seven items rated on a 4-point Likert scale (0 = "Not at all" to 3 = "Very much"). Responses were summed to compute an overall anxiety score. Participants scoring below 10 were labeled as "non-anxious," while those scoring 10 or higher were as "anxious," following validated DASS-21 thresholds.

After the study, PPG, EDA, and the recorded video of the Speech task were extracted for further analysis. A total of 108 (60 anxious and 48 non-anxious) participants took part in our study, including 81 males and 27 females. The average age of the participants was 19.73 (±1.34). As a token of appreciation for their time, they received a refreshment box upon completing the study.

Social Anxiety Dataset (SAD): To evaluate the robustness HCFSLN, we tested it on Social Anxiety Dataset (SAD) [31, 33, 34], which includes all modalities present in M2AD. In the SAD study, participants performed a 2.5-minute anxiety provoking speech activity in front of an audience while their ECG, PPG, and EDA signals were passively recorded. Additionally, the speech activity was captured using an audiovisual camera. We used data from 92 participants (41 anxious and 51 non-anxious) to benchmark our framework for anxiety detection.

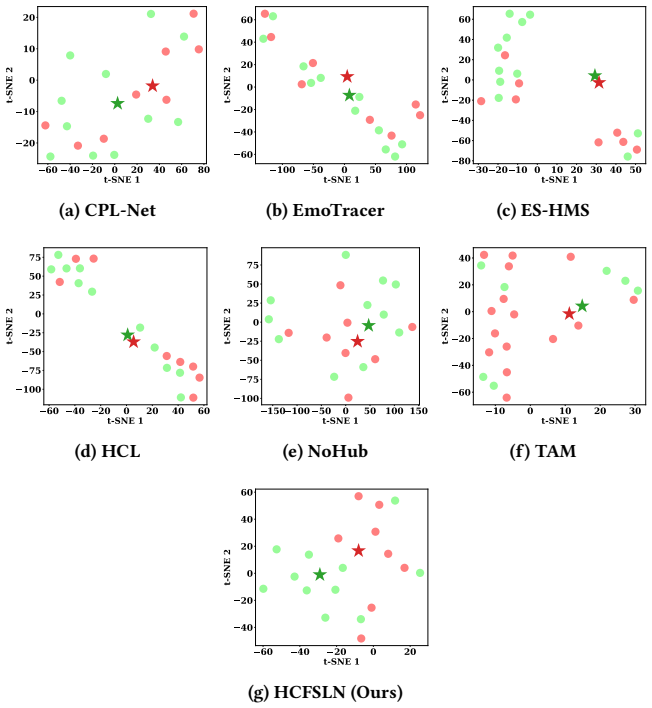
4.2 Implementation Details

Data preprocessing: Though the speech activity in **M2AD** was of 120 seconds, some instances had slight variations of approximately (±) 5 seconds in the data. To address this, we standardized the data length to 120 seconds. If any data modality exceeded 120 seconds, we trimmed it to the first 120 seconds, and if the data length was shorter than 120 seconds, we padded it with zeros to make it up to 120 seconds. For feature extraction, we represented each participant's recording as a sequence of non-overlapping 1-second segments, resulting in a fixed temporal length of 120 steps. The resulting 1 Hz representation refers to the temporal resolution of the final feature sequence, not naive downsampling of the raw signals by retaining a single sample per second. For audio, which was recorded at 44.1 kHz, acoustic and cepstral descriptors were computed from the original signal within each 1-second segment using Librosa, producing a segment-level feature vector that preserves spectral and temporal information from the full window. We did not use transcript-based text features because linguistic content is highly dependent on the assigned speech topic and may introduce topic-specific bias rather than capturing participants' anxiety characteristics [34]. Therefore, we focused on topic-independent behavioral and physiological modalities (audio, video, PPG, and EDA). For the video modality, recorded at approximately 27 frames per second, we used OpenFace[2] to extract 711 features related to eye and head position, face landmarks, and facial action units from all frames within each 1-second segment, followed by temporal aggregation [1]. In audio, we followed the methodology of Sahu et al. and Mohammadi et al. to extract 114 audio features using the Librosa Python package[3]; the complete audio feature list is provided in the supplementary material [28, 34]. For the PPG signal, we first cleaned the noisy data using the NeuroKit Python package[4], then summarized the cleaned 512 Hz signal over each 1-second segment to match the temporal resolution of the audio and video feature sequences [23]. Similarly, the EDA signal was cleaned using a Butterworth filter from NeuroKit and decomposed into phasic and tonic components, after which segment-level summaries were computed over each 1-second window. The **SAD** dataset was processed in a similar manner as mentioned above. The extracted features are concatenated into a shape (N, L, D), where N represents the number of participants, L is the fixed sequence length of 120, and D is the total feature dimension across modalities. This structure preserves temporal dependencies while enabling efficient sequence-based learning [24].

4.3 Baseline Models

We benchmarked HCFSLN against **six** recent SOTA FSL networks: Few-shot Sequence Learning with Transformers (TAM) [20]—adapted

[2]https://github.com/TadasBaltrusaitis/OpenFace
[3]https://librosa.org/doc/latest/index.html
[4]https://github.com/neuropsychology/NeuroKit



(a) CPL-Net (b) EmoTracer (c) ES-HMS

(d) HCL (e) NoHub (f) TAM

(g) HCFSLN (Ours)

Figure 3: t-SNE prototype visualizations of models.

with LSTM encoders for temporal sequence data; Uniform Hyperspherical Structure-preserving Embeddings (noHub) [37]; Hyperbolic Few-Shot Contrastive Learning (HCL) [3]—using LSTM encoders; Efficient Spatiotemporal Few-Shot Learning in Hyperbolic Space (ES-HMS) [41]—implemented with LSTM modality encoders and hyperbolic prototype fusion; EmoTracer [18]—LSTM-based fusion; and Contrastive Prototypical Learning (CPL-Net) [39]—using transformer-based temporal attention for sequential data. These SOTA baselines were selected to represent the major FSL approaches using Euclidean, hyperspherical, and hyperbolic embeddings. We trained and evaluated each method under both one-shot and five-shot settings to assess performance across varying sample sizes.

Few-Shot Training Configuration: StandardScaler normalization, sequence padding, and stratified train-test splitting (*test_size* = 0.2) are applied across all models. Training is conducted using the Adam optimizer (*learning rate* = $1e^{-3}$, 50 epochs), with each episode selecting K support samples ($K \in \{1, 5\}$) per class (anixous, non-anxious) . In the proposed FSL framework, a residual hyperbolic block with an adaptive projection layer (trainable curvature initialized as $\alpha = 1.0$) refines representations. Optimization combines hyperbolic prototypical loss and angular margin loss ($\gamma = 0.2$, $\lambda = 1.0$) to enhance class separability. Each experiment is repeated 5 times, reporting average accuracy with standard deviations [10] for model stability and robustness. *All experiments were conducted on a system equipped with an NVIDIA RTX 4070 Ti 12GB GPU, 64 GB RAM, Win 11 and an Intel i9 12th Gen CPU.*

4.4 Results

Table 1 highlights clear, quantifiable trends across models, modalities, and shot settings on the SAD and M2AD datasets, presenting the best performing modality combinations. HCFSLN consistently outperforms all baselines across datasets, modalities, and shot settings, proving its effectiveness for few-shot anxiety detection. In single-modality settings under low-shot (1-shot), audio (A) performs best, achieving 88% accuracy on SAD and 79% on M2AD, highlighting its strong predictive power as a non-intrusive behavioral modality. For dual modalities, combinations involving audio and physiological signals such as Audio-EDA (A-E) and Audio-PPG (A-P) improve accuracy by approximately 3–5% over unimodal audio, demonstrating the complementary value of adding intrusive physiological modalities captured via wearable sensors. Tri-modal fusion, combining audio, physiological, and video data (e.g., A-P-V or A-E-V), achieves the highest accuracies—up to 86% on SAD and 83% on M2AD—outperforming unimodal and dual setups by 5–7%, underscoring the advantage of leveraging multiple data sources. Interestingly, increasing from 1-shot to 5-shot settings generally maintains or slightly improves performance, with tri-modal fusion showing the most stable or improved accuracy, while some unimodal and dual-modal results experience minor decreases (around 4–6%), likely due to data complexity. Also using all modality often reduce accuracy due to addition of extra noise and data complexity. Across all modalities, HCFSLN exhibits significantly lower variance, reflecting more consistent and reliable learning compared to other models.

Table 2 reports macro-F1 scores to assess class-balanced performance across anxious and non-anxious participants. HCFSLN achieves the strongest macro-F1 results across all modality groups on both datasets, indicating that its improvements are not driven only by the majority class. On SAD, HCFSLN reaches 0.91 with A-E in the 1-shot setting and 0.86 with audio alone in the 5-shot setting, showing reliable classification even when only a few labeled examples are available. On M2AD, HCFSLN obtains 0.85 with A-V in the 1-shot setting and 0.88 using all modalities in the 5-shot setting, suggesting that the adaptive fusion mechanism helps preserve class-level information under multimodal settings. Overall, the macro-F1 results strengthen the accuracy findings by showing that HCFSLN improves balanced anxiety detection rather than only increasing aggregate performance.

Overall non-intrusive modalities (audio, video) offer a strong baseline, while adding intrusive physiological signals significantly boosts accuracy—especially in multimodal fusion—highlighting the importance of modality choice based on data availability and practical constraints.

5 Prototype Visualizations

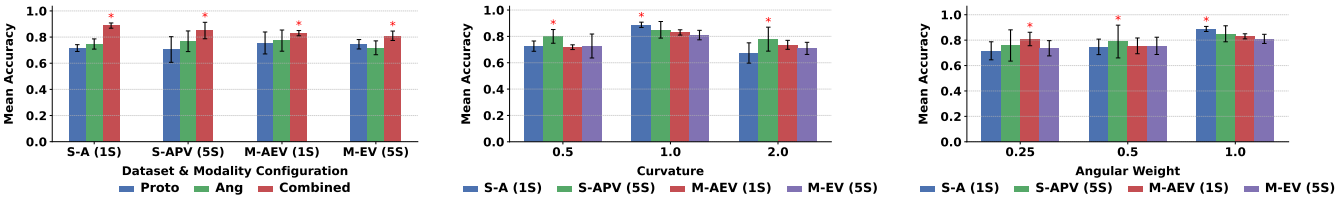
Figure 3 presents the best prototype t-SNE visualizations for each model, selected from five runs based on highest overall accuracy to ensure a fair comparison. The plots illustrate clusters centered on their prototypes. HCFSLN exhibits particularly compact clusters with centrally located prototypes, reflecting strong discriminative ability. In contrast, models using Euclidean embeddings (EmoTracer, TAM, CPL-Net) display more dispersed and overlapping clusters,

Model	Shots	SAD				M2AD			
		Single	Dual	Triple	All	Single	Dual	Triple	All
TAM	1	A(0.63, 0.05)	E-P(0.66, 0.03)	E-P-V(0.63, 0.07)	(0.61, 0.03)	V(0.69, 0.06)	P-V(0.74, 0.04)	E-P-V(0.69, 0.10)	(0.66, 0.07)
	5	P(0.66, 0.03)	E-V(0.66, 0.06)	A-E-P(0.67, 0.04)	(0.64, 0.06)	V(0.66, 0.14)	P-V(0.67, 0.13)	E-P-V(0.71, 0.11)	(0.63, 0.08)
NoHub	1	A(0.71, 0.03)	A-E(0.67, 0.04)	A-E-P(0.65, 0.08)	(0.55, 0.13)	V(0.66, 0.10)	A-E(0.68, 0.07)	E-P-V(0.66, 0.11)	(0.66, 0.05)
	5	A(0.70, 0.10)	A-E(0.65, 0.06)	E-P-V(0.55, 0.13)	(0.52, 0.04)	V(0.70, 0.11)	A-E(0.66, 0.13)	A-E-P(0.66, 0.12)	(0.65, 0.08)
HCL	1	V(0.73, 0.06)	E-P(0.68, 0.07)	A-E-P(0.71, 0.08)	(0.66, 0.09)	V(0.65, 0.06)	P-V(0.68, 0.07)	A-P-V(0.67, 0.13)	(0.57, 0.05)
	5	P(0.71, 0.05)	A-P(0.67, 0.07)	E-P-V(0.63, 0.06)	(0.60, 0.06)	E(0.69, 0.11)	E-P(0.68, 0.09)	A-E-P(0.65, 0.05)	(0.58, 0.07)
ES-HMS	1	P(0.68, 0.00)	E-P(0.72, 0.05)	E-P-V(0.65, 0.06)	(0.66, 0.06)	V(0.64, 0.05)	P-V(0.70, 0.06)	A-E-P(0.67, 0.09)	(0.63, 0.07)
	5	P(0.68, 0.04)	E-P(0.66, 0.13)	A-E-V(0.63, 0.08)	(0.63, 0.05)	E(0.69, 0.06)	E-P(0.66, 0.07)	E-P-V(0.64, 0.07)	(0.57, 0.06)
EmoTracer	1	P(0.67, 0.06)	E-P(0.66, 0.04)	E-P-V(0.62, 0.10)	(0.67, 0.10)	E(0.65, 0.06)	E-P(0.62, 0.12)	A-E-P(0.60, 0.09)	(0.60, 0.12)
	5	P(0.71, 0.10)	E-P(0.66, 0.10)	A-P-V(0.60, 0.06)	(0.58, 0.10)	P(0.65, 0.14)	A-P(0.61, 0.06)	A-E-V(0.63, 0.05)	(0.65, 0.13)
CPL-Net	1	P(0.70, 0.04)	P-V(0.70, 0.04)	E-P-V(0.72, 0.08)	(0.63, 0.08)	E(0.59, 0.08)	A-P(0.62, 0.12)	A-P-V(0.62, 0.11)	(0.56, 0.05)
	5	V(0.66, 0.08)	A-P(0.65, 0.03)	A-P-V(0.67, 0.06)	(0.62, 0.04)	E(0.60, 0.05)	A-P(0.66, 0.05)	A-P-V(0.67, 0.06)	(0.62, 0.06)
HCFSLN (Ours)	1	**A(0.88, 0.02)**	**A-E(0.85, 0.07)**	**A-P-V(0.85, 0.06)**	**(0.79, 0.07)**	**A(0.79, 0.07)**	A-**P(0.82, 0.08)**	**A-E-V(0.83, 0.02)**	**(0.75, 0.05)**
	5	**A(0.82, 0.07)**	**A-E(0.82, 0.04)**	**A-P-V(0.86, 0.05)**	**(0.80, 0.05)**	**A(0.78, 0.06)**	**E-V(0.81, 0.03)**	**A-E-V(0.79, 0.09)**	**(0.78, 0.03)**

Table 1: Performance comparison of models across modalities (single, dual, triple, and all) and shot settings (1-shot and 5-shot). Modalities—Audio, Video, PPG, and EDA—are abbreviated as A, V, P, and E, respectively. For each model, only the best-performing modality combination(s) are shown under 1 and 5-shot settings.

Model	Shots	SAD				M2AD			
		Single	Dual	Triple	All	Single	Dual	Triple	All
TAM	1	P(0.66, 0.05)	A-V(0.65, 0.06)	A-P-V(0.66, 0.05)	(0.63, 0.06)	V(0.72, 0.09)	E-V(0.76, 0.10)	A-P-V(0.74, 0.11)	(0.67, 0.08)
	5	P(0.64, 0.05)	E-P(0.66, 0.07)	E-P-V(0.60, 0.08)	(0.62, 0.09)	V(0.69, 0.07)	E-V(0.72, 0.13)	E-P-V(0.65, 0.07)	(0.66, 0.13)
NoHub	1	P(0.66, 0.05)	A-E(0.66, 0.10)	A-E-P(0.56, 0.12)	(0.52, 0.09)	A(0.64, 0.03)	P-V(0.64, 0.07)	A-P-V(0.66, 0.08)	(0.62, 0.04)
	5	A(0.67, 0.09)	A-E(0.69, 0.02)	A-E-P(0.62, 0.08)	(0.54, 0.04)	V(0.66, 0.05)	A-E(0.65, 0.14)	A-P-V(0.67, 0.09)	(0.56, 0.11)
HCL	1	P(0.68, 0.04)	A-P(0.67, 0.09)	E-P-V(0.67, 0.07)	(0.66, 0.07)	V(0.68, 0.07)	E-V(0.69, 0.07)	A-P-V(0.63, 0.04)	(0.59, 0.05)
	5	P(0.71, 0.05)	E-P(0.66, 0.05)	A-E-P(0.66, 0.12)	(0.59, 0.02)	E(0.65, 0.07)	E-P(0.67, 0.03)	E-P-V(0.65, 0.10)	(0.58, 0.07)
ES-HMS	1	P(0.71, 0.08)	P-V(0.74, 0.12)	E-P-V(0.69, 0.04)	(0.65, 0.03)	E(0.68, 0.05)	E-P(0.69, 0.08)	A-E-P(0.70, 0.10)	(0.62, 0.05)
	5	P(0.69, 0.02)	E-P(0.68, 0.06)	A-E-V(0.64, 0.07)	(0.66, 0.05)	E(0.71, 0.02)	E-P(0.76, 0.06)	E-P-V(0.61, 0.02)	(0.57, 0.06)
EmoTracer	1	P(0.68, 0.08)	E-P(0.61, 0.16)	E-P-V(0.58, 0.10)	(0.61, 0.05)	E(0.65, 0.13)	A-V(0.59, 0.07)	A-E-V(0.58, 0.10)	(0.58, 0.12)
	5	P(0.67, 0.07)	E-V(0.60, 0.05)	A-E-V(0.55, 0.08)	(0.52, 0.07)	E(0.68, 0.10)	A-E(0.69, 0.07)	A-E-V(0.60, 0.08)	(0.61, 0.09)
CPL-Net	1	P(0.71, 0.06)	P-V(0.72, 0.12)	E-P-V(0.71, 0.06)	(0.64, 0.04)	P(0.65, 0.10)	A-E(0.66, 0.04)	A-E-P(0.65, 0.06)	(0.57, 0.08)
	5	A(0.65, 0.03)	E-V(0.71, 0.07)	A-E-P(0.67, 0.04)	(0.60, 0.06)	A(0.71, 0.07)	A-E(0.73, 0.14)	E-P-V(0.69, 0.07)	(0.69, 0.07)
HCFSLN (Ours)	1	**V(0.79, 0.04)**	**A-E(0.91, 0.09)**	**A-E-V(0.84, 0.06)**	**(0.87, 0.06)**	**A(0.78, 0.06)**	**A-V(0.85, 0.04)**	**A-E-V(0.79, 0.05)**	**(0.78, 0.14)**
	5	**A(0.86, 0.05)**	**A-P(0.84, 0.09)**	**A-E-V(0.81, 0.06)**	**(0.83, 0.04)**	**A(0.75, 0.09)**	**A-V(0.81, 0.06)**	**A-E-V(0.84, 0.06)**	**(0.88, 0.08)**

Table 2: Macro-F1 score comparison of models across modalities (single, dual, triple, and all) and shot settings (1-shot and 5-shot). Modalities—Audio, Video, PPG, and EDA—are abbreviated as A, V, P, and E, respectively. For each model, only the best-performing modality combination(s) are shown under 1 and 5-shot settings.



(a) Impact of loss type combinations. **(b) Impact of varying curvature value.** **(c) Impact of varying angular weights λ_{ang}**

Figure 4: Ablation results showing the effect of loss type, curvature, and angular weight on accuracy.

while hyperspherical embeddings (NoHub) show moderate clustering but less distinct class separation. Hyperbolic embeddings (ES-HMS, HCL, HCFSLN) effectively capture complex data relationships, as evidenced by their tighter, well-separated clusters, highlighting their superior representational power for few-shot anxiety detection.

6 Ablation Studies

We evaluated how loss functions, angular margin weight (λ), and hyperbolic curvature (α) affect HCFSLN accuracy in 1-shot and 5-shot settings for the best configurations. We picked best configuration of HCFSLN in both SAD and M2AD datasets as SAD-Audio (S-A) (1 shot), SAD Audio-PPG-Video (S-APV) (5 shot), M2AD Audio-EDV-Video (M-AEV) (1 shot), and M2AD EDA-Video (M-EV) (5 shot). We performed Welch's t-tests [35] to evaluate the significance of all three ablation loss types, as well as the effects of hyperbolic curvature and angular weights.

Impact of Loss Type: Fig. 4a compares the effects of Hyperbolic Prototypical Loss (Proto), Angular Loss (Ang), and their combination (Combined) on accuracy across datasets and modalities. The combined loss consistently outperforms individual losses. Specifically, (S-A) (1 shot) combined significantly surpasses both proto and angular losses ($p < 0.001$). For (M-AEV) (1 Shot), combined outperforms proto alone ($p = 0.036$), and in (M-EV) (5 shot), it significantly improves over angular loss ($p = 0.012$). No significant difference was observed in the (S-APV) (5 Shot) setting. Overall, combining hyperbolic and angular margin losses improves robustness and accuracy in few-shot settings.

Impact of Hyperbolic Curvature (α): We assessed the impact of curvature on HCFSLN by comparing accuracy means at $\alpha = 0.5, 1.0$, and 2.0 across datasets and modality configurations. As shown in Fig. 4b, for (S-A) (1-shot) accuracy increased by 17.5% ($p = 0.00007$) from (α) 0.5 to 1.0, then declined significantly at 2.0 compared to 1.0 ($p = 0.0026$). In (S-APV) (5-shot), accuracy peaked at $\alpha = 1.0$ with stable performance and no significant differences across curvatures ($p > 0.18$). For (M-AEV) (1-shot), accuracy improved by 15.5% from (α) 0.5 to 1.0 ($p = 0.001$), with a marginally non-significant change between (α) 1.0 and 2.0 ($p = 0.06$). The (M-EV) (5-shot) setting showed moderate, non-significant gains at 1.0 over 0.5 ($p = 0.24$). Overall, $\alpha = 1.0$ consistently achieves the highest and most robust accuracy, underscoring the crucial role of hyperbolic curvature in enhancing representations.

Imapact of Angular Weights: In (Fig. 4c) we analyze the effect of angular loss weight (λ) on accuracy . Significant accuracy improvements were observed when increasing λ from 0.25 to 1.0 in all settings, with gains up to 17.2% (e.g., $p = 0.0007$ for (S-A) (1-shot)), confirming the benefit of angular loss. Comparisons between 0.25 and 0.5, and 0.5 and 1.0 showed mixed effect (p ranging 0.02–0.25), indicating a gradual but performance increase as λ grows. These results shows that a combined loss with a higher angular weight optimally balances hyperbolic and angular components, enhancing accuracy in few-shot multimodal anxiety detection.

7 Limitations

The proposed framework has a few limitations. First, it relies on the availability and synchronization of multiple modalities (audio, physiological signals, and video), which may not always be feasible in real-world deployments. Second, variations in data quality across modalities due to noise, sensor differences, or environmental factors can influence model performance. Third, the multimodal fusion and hyperbolic representation introduce additional computational overhead, which may pose challenges for deployment on resource-constrained systems.

8 Safe and Responsible Innovation Statement

This work addresses multimodal anxiety detection, which involves sensitive behavioral and physiological data. We ensure responsible use by collecting data with informed consent and anonymizing all participant information to protect privacy. While our dataset includes diverse modalities, its limited size may introduce biases; future work will expand demographic diversity to improve fairness and generalizability. The proposed system is intended to support, not replace, clinical decision-making. Potential misuse, such as surveillance without consent, is a concern, and safeguards should be enforced in deployment. Overall, we emphasize privacy, transparency, and ethical integration of multimodal systems for safe and responsible mental health assessment.

9 Conclusion

In this work, we introduced HCFSLN for multimodal anxiety detection, addressing the challenge of small mental health datasets. By integrating speech, physiological signals (PPG, EDA), and video features, our framework leverages hyperbolic embeddings, cross-modal attention, and an adaptive gating network to enhance feature separability and generalization. Experiments show HCFSLN outperforms existing few-shot learning models. It can be integrated into wearables and smartphone apps for real-time anxiety assessment, enabling early detection and intervention. This benefits mental health professionals, workplaces, and remote healthcare by providing accessible, proactive anxiety monitoring.

References

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 1–10.

[2] Mehmet Baygin, Prabal Datta Barua, Sengul Dogan, Turker Tuncer, Tan Jen Hong, Sonja March, Ru-San Tan, Filippo Molinari, and U Rajendra Acharya. 2024. Automated anxiety detection using probabilistic binary pattern with ECG signals. *Computer Methods and Programs in Biomedicine* 247 (2024), 108076.

[3] Won June Choi, Jin Hwangbo, Quan Anh Duong, Jae-Hyeok Lee, and Jin Kyu Gahm. 2024. Differentiating atypical parkinsonian syndromes with hyperbolic few-shot contrastive learning. *NeuroImage* 304 (2024), 120940.

[4] Yassine Daadaa. 2024. EmotionNet: Dissecting Stress and Anxiety Through EEG-based Deep Learning Approaches. *International Journal of Advanced Computer Science & Applications* 15, 1 (2024).

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.

[6] Ning Ding, Yulin Chen, Ganqu Cui, Xiaobin Wang, Hai-Tao Zheng, Zhiyuan Liu, and Pengjun Xie. 2022. Few-shot classification with hypersphere modeling of prototypes. *arXiv preprint arXiv:2211.05319* (2022).

[7] Alae Eddine El Hmimdi and Zoï Kapoula. 2024. Physiological Data Augmentation for Eye Movement Gaze in Deep Learning. *BioMedInformatics* 4, 2 (2024), 1457–1479.

[8] Kexin Feng and Theodora Chaspari. 2021. Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing* 14, 2 (2021), 1627–1633.

[9] Kexin Feng, Jacqueline B Duong, Kayla E Carta, Sierra Walters, Gayla Margolin, Adela C Timmons, and Theodora Chaspari. 2022. A few-shot learning approach with domain adaptation for personalized real-life stress detection in close relationships. *arXiv preprint arXiv:2210.15247* (2022).

[10] Minghao Fu, Yun-Hao Cao, and Jianxin Wu. 2022. Worst case matters for few-shot recognition. In *European Conference on Computer Vision*. Springer, 99–115.

[11] Mina Ghadimi Atigh, Martin Keller-Ressel, and Pascal Mettes. 2021. Hyperbolic busemann learning with ideal prototypes. *Advances in neural information processing systems* 34 (2021), 103–115.

[12] Hao Guo, Jiuyang Tang, Weixin Zeng, Xiang Zhao, and Li Liu. 2021. Multi-modal entity alignment in hyperbolic space. *Neurocomputing* 461 (2021), 598–607.

[13] Md Monirul Islam, Shahriar Hassan, Sharmin Akter, Ferdaus Anam Jibon, and Md Sahidullah. 2024. A comprehensive review of predictive analytics models for mental illness using machine learning algorithms. *Healthcare Analytics* (2024), 100350.

[14] Syed Fahad Javaid, Ibrahim Jawad Hashim, Muhammad Jawad Hashim, Emmanuel Stip, Mohammed Abdul Samad, and Alia Al Ahbabi. 2023. Epidemiology of anxiety disorders: global burden and sociodemographic associations. *Middle East Current Psychiatry* 30, 1 (2023), 44.

[15] Chandan Kumar Jha and Maheshkumar H Kolekar. 2020. Cardiac arrhythmia classification using tunable Q-wavelet transform based features and support vector machine classifier. *Biomedical Signal Processing and Control* 59 (2020), 101875.

[16] Xiaogang Jia, Yusong Tan, Songlei Jian, and Yonggang Che. 2023. Multi-scale Recurrent LSTM and Transformer Network for Depth Completion. *arXiv preprint arXiv:2309.16301* (2023).

[17] Clemens Kirschbaum, Karl-Martin Pirke, and Dirk H Hellhammer. 1993. The 'Trier Social Stress Test'–a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology* 28, 1-2 (1993), 76–81.

[18] Wenting Kuang, Shuo Jin, Danyang Wang, Yuda Zheng, Yongpan Zou, and Kaishun Wu. 2024. EmoTracer: A User-independent Wearable Emotion Tracer with Multi-source Physiological Sensors Based on Few-shot Learning. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2680–2687.

[19] Lingxiao Li, Yi Zhang, and Shuhui Wang. 2023. The euclidean space is evil: Hyperbolic attribute editing for few-shot image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 22714–22724.

[20] Lajanugen Logeswaran, Ann Lee, Myle Ott, Honglak Lee, Marc'Aurelio Ranzato, and Arthur Szlam. 2020. Few-shot sequence learning with transformers. *arXiv preprint arXiv:2012.09543* (2020).

[21] SH Lovibond. 1995. Manual for the depression anxiety stress scales. *Psychology Foundation of Australia* (1995).

[22] Rongkai Ma, Pengfei Fang, Tom Drummond, and Mehrtash Harandi. 2022. Adaptive poincaré point to set distance for few-shot classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 1926–1934.

[23] Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and SH Annabel Chen. 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior research methods* (2021), 1–8.

[24] Sreevalsan S Menon and K Krishnamurthy. 2021. Multimodal ensemble deep learning to predict disruptive behavior disorders in children. *Frontiers in neuroinformatics* 15 (2021), 742807.

[25] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. 2024. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision* 132, 9 (2024), 3484–3508.

[26] Haimiao Mo, Siu Cheung Hui, Xiao Liao, Yuchen Li, Wei Zhang, and Shuai Ding. 2024. A multimodal data-driven framework for anxiety screening. *IEEE Transactions on Instrumentation and Measurement* (2024).

[27] Haimiao Mo, Yuchen Li, Shanlin Yang, Wei Zhang, and Shuai Ding. 2022. SFF-DA: Sptialtemporal Feature Fusion for Detecting Anxiety Nonintrusively. *arXiv preprint arXiv:2208.06411* (2022).

[28] Mohsen Mohammadi and Hamid Reza Sadegh Mohammadi. 2017. Robust features fusion for text independent speaker verification enhancement in noisy environments. In *2017 Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 1863–1868.

[29] Gabriel Moreira, Manuel Marques, João Paulo Costeira, and Alexander Hauptmann. 2024. Hyperbolic vs Euclidean embeddings in few-shot learning: Two sides of the same coin. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2082–2090.

[30] Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems* 30 (2017).

[31] Nilesh Kumar Sahu, Snehil Gupta, and Haroon Lone. 2024. Wearable Technology Insights: Unveiling Physiological Responses During Three Different Socially Anxious Activities. *ACM Journal on Computing and Sustainable Societies* (2024).

[32] Nilesh Kumar Sahu, Snehil Gupta, and Haroon R Lone. 2025. MAD: A Multimodal Physiological and Self-Reported Dataset for Anxiety Research from a Low-to-Middle-Income Country. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 4 (2025), 1–34.

[33] 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).

[34] Nilesh Kumar Sahu, Manjeet Yadav, and Haroon R Lone. 2024. Unveiling Social Anxiety: Analyzing Acoustic and Linguistic Traits in Impromptu Speech within a Controlled Study. *ACM Journal on Computing and Sustainable Societies* (2024).

[35] Tetsuya Sakai. 2016. Two sample t-tests for ir evaluation: Student or welch?. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 1045–1048.

[36] Ermal Toto, ML Tlachac, and Elke A Rundensteiner. 2021. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM international conference on information & knowledge management*. 4145–4154.

[37] Daniel J Trosten, Rwiddhi Chakraborty, Sigurd Løkse, Kristoffer Knutsen Wickstrøm, Robert Jenssen, and Michael C Kampffmeyer. 2023. Hubs and hyperspheres: Reducing hubness and improving transductive few-shot learning with hyperspherical embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7527–7536.

[38] Matteo Vismara, Nicolaja Girone, Giovanna Cirnigliaro, Federica Fasciana, Simone Vanzetto, Luca Ferrara, Alberto Priori, Claudio D'Addario, Caterina Viganò, and Bernardo Dell'Osso. 2020. Peripheral biomarkers in DSM-5 anxiety disorders: an updated overview. *Brain Sciences* 10, 8 (2020), 564.

[39] Liangguo Wang and Yiping Wang. 2025. A Multi-Modal Few-Shot Learning Framework for Depression Detection Using Video and Audio Features. In *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*. IEEE, 2519–2522.

[40] Xin Wang, Lei Cao, Huijun Zhang, Ling Feng, Yang Ding, and Ningyun Li. 2022. A meta-learning based stress category detection framework on social media. In *Proceedings of the ACM Web Conference 2022*. 2925–2935.

[41] Yifan Wei and Xueyun Nie. 2024. ES-HMS: Efficient Spatiotemporal Few-Shot Learning in Hyperbolic Space. In *Proceedings of the International Conference on Algorithms, Software Engineering, and Network Security*. 493–500.

[42] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*. PMLR, 10524–10533.

[43] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12226–12234.

[44] Kirsten Zantvoort, Barbara Nacke, Dennis Görlich, Silvan Hornstein, Corinna Jacobi, and Burkhardt Funk. 2024. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *npj Digital Medicine* 7, 1 (2024), 361.

[45] Haiping Zhang, Zepeng Hu, Dongjin Yu, Liming Guan, Xu Liu, and Conghao Ma. 2024. Multipath Attention and Adaptive Gating Network for Video Action Recognition. *Neural Processing Letters* 56, 2 (2024), 124.