https://doi.org/10.3311/PPee.42499

Periodica Polytechnica Electrical Engineering and Computer Science

Multimodal Deep Learning Driven English-speaking Emotion Recognition and Adaptive Teaching Strategy Generation

Yuebin Wang[1*]

[1] Department of Tourist Management, Henan Vocational College of Agriculture, Zhengzhou, 450000 Henan, China
[*] Corresponding author, e-mail: g29993195taolia@163.com

Received: 14 October 2025, Accepted: 02 June 2026, Published online: 09 July 2026

Abstract

It is critical for emotion-aware multichannel adaptation techniques to have high accuracy in terms of recognizing emotions, particularly in English-speaking learners' cases. Recurrent neural network-based methods, aggregation-based approaches, and conventional multimodal fusion techniques are known to have some drawbacks related to time dependency, dynamicity of emotional transitions, and inter-speaker variance robustness. In an attempt to address these limitations, we propose a task-specific multimodal framework that employs an emotion-structured transformer encoder coupled with a progressive emotion distillation strategy. What should be noted first about this paper is that its contribution lies in neither technique employed, since they were previously utilized in different works. The key idea behind the proposed approach is the combination of existing emotion recognition techniques into a multimodal pipeline that facilitates improved emotion recognition and robust representation learning. Estimation of the speaker-independent representation, emotion-aware representation, and improved emotion recognition are achieved through employing emotion-prior masked attention, emotion-gated feature transformation, multi-head attention pooling, and boundary-aware auxiliary supervision in ESTE. Further improvement in training the recognition model is obtained by incorporating emotionally ambiguous samples progressively and a distillation process based on soft labels, in addition to auxiliary tasks such as valence-arousal regression and boundary detection. Improved results are empirically observed when compared with selected baselines using emotion recognition benchmark datasets. As a conclusion, an interpretative teaching strategy generation method relying on the rules of teaching guidelines is suggested.

Keywords

emotion recognition, transformer encoder, curriculum learning, speech processing, affective computing

1 Introduction

The task of recognizing emotions in English-speaking and generating adaptive teaching strategies is increasingly critical in enhancing personalized education and emotional interaction systems [1]. Accurate emotion recognition improves communication efficiency and learning engagement, and also enables adaptive interventions that cater to individual learner needs [2]. Traditional education strategies often lack emotional sensitivity, which can hinder student motivation and performance [3]. Moreover, with the global adoption of online and remote learning, there is a rising demand for systems that can sense learners' emotional states and dynamically adjust teaching methods [4]. Hence, research into multimodal deep learning for emotion recognition and adaptive strategy generation is not only necessary but also timely, offering potential for more humane, intelligent, and effective education [5].

Early systems addressing emotional interaction in educational settings were constructed around rigid rule systems derived from psychological theory and linguistic observations [6]. These models attempted to define emotional categories by encoding logical mappings between speech prosody, word choice, and facial expressions. While these techniques facilitated controlled experimentation and human-interpretable outputs, they were limited in handling spontaneous, ambiguous, or culturally diverse expressions [7]. Variability in pronunciation, speaking pace, and expressive range further undermined their robustness in real-world teaching scenarios. These shortcomings prompted researchers to shift toward approaches that could better accommodate expressive variation and adapt to the fluidity of learner responses [8].

In response, a new wave of algorithms emerged that relied on statistical patterns observed in large datasets of human

Cite this article as: Wang, Y. "Multimodal Deep Learning Driven English-speaking Emotion Recognition and Adaptive Teaching Strategy Generation", Periodica Polytechnica Electrical Engineering and Computer Science, 2026. https://doi.org/10.3311/PPee.42499

speech and interaction [9]. These models leveraged feature extraction pipelines to quantify affective cues, such as pitch modulation, articulation rate, or visual micro expressions, and linked them to emotion labels through discriminative learning models [10]. Systems trained with this approach showed improved flexibility across speaking styles and better tolerance to noise [11]. Nonetheless, they still needed careful feature design and struggled to model complex temporal dependencies or cross-modal coherence [12]. Moreover, their scalability was often limited when faced with the multimodal richness of educational interactions involving spoken, textual, and visual cues.

Recent advances in neural computation have introduced configurations capable of learning expressive representations directly from raw input streams, significantly enhancing emotion modeling fidelity [13]. Deep learning (DL) models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), combined with attention mechanisms, have demonstrated strong performance in capturing emotion dynamics within audio and facial expression sequences [14]. More recently, transformer-based models and pre-trained backbones, like ERT for textual inputs or wav2vec for speech, have enabled fine-grained contextual learning and transfer across domains [15]. Multimodal fusion techniques have further strengthened emotion inference by aligning and integrating diverse modalities within unified representations. Despite these successes, current systems still face challenges such as modality imbalance, resource-intensive training, and the need for pedagogically grounded adaptation mechanisms [16]. Continued research is therefore required to design lightweight, interpretable, and pedagogically adaptive models for real-time deployment in emotionally responsive learning environments.

Due to the limitations of rule-based systems, classical machine learning algorithms, and recent deep learning architectures, we propose a multimodal approach to emotion recognition and teaching adaptation in English-speaking settings. This paper does not intend to create a completely new architecture for the Transformer model and curriculum learning strategy. Rather, its merit lies in bringing together several ideas, including emotion-aware temporal features, gate-controlled feature operations, utterance-level summarization using an attention mechanism, distillation on an increasing dataset, and teaching decision mapping. The goal of this paper is to increase the correlation of the outputs from emotion recognition and teaching adaptation processes in English-speaking learning environments.

The most important contributions made by the paper are as follows:

- Emotion-Structured Transformer Encoder, which can be used specifically for the task of emotion recognition in the English language. This network architecture employs emotion-prior masked self-attention, emotion-gated feedforward transformation, multi-head attention fusion, and speaker identity normalization.
- The use of the Progressive Emotion Distillation Strategy allows addressing the challenge of training data due to emotional ambiguity. It includes curriculum learning, knowledge distillation using soft labels, sampling with uncertainties, and supplementary supervision via valence/arousal regression and emotion boundary detection.
- Teaching strategy generator, which is capable of generating an emotion-based adaptive teaching strategy based on the output of the emotion recognizer. This approach aims to assign emotions to appropriate pedagogically relevant strategies, such as making the task easier if frustration is detected, reducing the complexity of the task if anxiety is experienced, etc.
- Detailed experiments, ablation tests, robustness analysis, and hypothesis testing have been performed to study the impact of each component.

2 Related work

2.1 Multimodal features for emotion recognition

Multimodal DL is a transformative approach for emotion recognition in English-speaking, combining multiple data sources such as audio, visual, and textual modalities to capture the nuanced expressions of emotion that a single modality might miss [17]. In traditional systems, speech emotion detection often relied solely on acoustic features, including pitch, energy, and spectral properties; however, these features alone frequently fall short in accurately representing complex emotional states. Multimodal frameworks integrate synchronized streams of audio signals, facial expressions, and linguistic cues, providing a richer and more holistic representation of speaker emotion [18]. Research has displayed that using CNNs for visual feature extraction and RNNs, including Long Short-Term Memory (LSTM) units for sequential modeling of audio signals, significantly enhances recognition accuracy. Attention mechanisms are frequently employed to weigh the significance of diverse modalities dynamically, allowing the scheme to adapt to the varying significance of speech tone versus facial movements depending on context [19]. Multimodal fusion strategies, whether early

fusion at the feature level, late fusion at the decision level, or hybrid fusion methods, are crucial for achieving high-performance emotion recognition. Datasets like The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [20] and Multimodal EmotionLines Dataset (MELD) [21] have been extensively used to benchmark multimodal approaches, and their annotations in both categorical and dimensional emotion models provide a comprehensive training ground for DL systems [22]. Emerging trends include the use of transformers for cross-modal learning and graph neural networks (GNNs) to model inter-modal relationships more effectively [23]. The complexity of natural spoken language, with its inherent variability in prosody, syntax, and pragmatic elements, demands sophisticated feature extraction and fusion techniques to achieve robust emotion recognition [24]. Advances in pretraining large multimodal schemes on broad databases and fine-tuning on task-specific corpora are also pushing the frontier of what multimodal deep learning can accomplish in English-speaking emotion recognition [25]. Dolatabadi et al. [26] performed experiments regarding the detection of emotions using EEG signals and peripheral physiological parameters including GSR, BVP, respiration rate, skin temperature, EMG, and EOG. For this purpose, various linear and nonlinear feature extraction approaches including PNN, GRNN, SVM, and KNN were utilized to detect valence, arousal, dominance, and liking. The following results were obtained from the above-mentioned approaches: EEG performed better than peripheral physiological parameters for detecting valence, arousal, and dominance; however, peripheral physiological parameters performed better than EEG for liking detection.

2.2 Adaptive teaching based on emotions

Adaptive teaching strategies grounded in emotion recognition aim to dynamically adjust instructional content, pace, and pedagogical approaches based on the emotional states of English-speaking learners, leading to more personalized and effective education experiences [27]. Conventional teaching paradigms that adopt a uniform approach are increasingly insufficient for heterogeneous learner demographics, especially in contexts of language acquisition where factors such as motivation, anxiety, and levels of confidence exert a substantial influence on performance. The incorporation of real-time emotion recognition technologies within computer-assisted language learning (CALL) settings enables educators to identify variations in students' emotional conditions and modify their interventions accordingly [28]. For example, should the system identify increased frustration *via* vocal stress indicators and negative facial expressions, it may autonomously simplify tasks, provide motivational support, or recommend a brief respite to mitigate cognitive strain. In contrast, indications of boredom could lead the system to present more demanding content or interactive experiences to re-engage the learner [29]. Researchers are investigating reinforcement learning algorithms to enhance adaptive methodologies, wherein the instructional agent discerns the most effective pedagogical actions based on ongoing emotional responses from students [30]. Furthermore, the implementation of explainable AI (XAI) techniques is crucial to ensure that the decisions made regarding adaptations are transparent and comprehensible to both instructors and learners. The amalgamation of affective computing into intelligent tutoring systems (ITS) necessitates advanced emotional modeling that considers not only ephemeral emotions but also enduring affective characteristics such as learner engagement and perseverance [31]. Multimodal emotion recognition improves these schemes by delivering more dependable and nuanced emotional data than unimodal approaches. Challenges persist in addressing cultural and individual differences in emotional expression, which underscores the necessity for the creation of culturally aware and personalized emotion-adaptive schemes. Ultimately, emotion-driven adaptive instruction signifies a notable progression in personalized education, possessing the potential to enhance learner outcomes, bolster motivation, and diminish attrition rates in English language learning initiatives [32].

2.3 Challenges and future research directions

Despite the positive advances noted in multimodal DL for emotion recognition in English speech and the derivation of adaptive learning approaches, several key barriers remain that require critical analysis in future studies [33]. A key challenge is the sparse availability of comprehensive, high-quality multimodal databases explicitly engineered for scenarios of English language learning, hindering the derivation of schemes that can generalize rigorously across diversified learner populations. The acquisition of such databases requires careful attention to issues of privacy, consent procedures, and demographic coverage [34]. Another pressing issue is the real-time processing requirements of adaptive systems, where latency in emotion detection and response generation can undermine the effectiveness of interventions. Lightweight schemes and efficient inference techniques are, therefore, crucial areas of exploration [35]. Moreover, current emotion recognition systems often struggle with ambiguity and mixed emotions, which are prevalent in naturalistic

speaking scenarios; future schemes must incorporate mechanisms for handling emotional complexity and uncertainty. Interpretability and fairness also pose significant challenges, as black-box deep learning schemes can inadvertently reinforce biases or make adaptation decisions that are difficult to justify to users [36]. Ethical considerations surrounding the surveillance and analysis of emotional states in educational settings must be thoroughly addressed, including transparent communication of data usage policies and the provision of opt-out options for learners. Cross-cultural generalizability is another vital area, as emotional expressions and their interpretations can vary significantly across cultural backgrounds; thus, schemes must be trained and validated on culturally diverse databases. Future research directions may involve the integration of self-supervised learning to better leverage unlabeled multimodal data, the development of zero-shot or few-shot learning capabilities to adapt to new learners with minimal supervision, and the incorporation of human-in-the-loop systems where teachers can guide or override model decisions based on pedagogical expertise. Exploring the synergy between emotion recognition and other learner state detection methods, such as cognitive load and attention monitoring, could also lead to more holistic and effective adaptive teaching systems [37].

3 Method

3.1 Overview

Emotion recognition from English speech is a fundamental problem that has attracted significant attention due to its wide range of applications, including human-computer interaction, mental health monitoring, and intelligent dialogue systems. Unlike textual or visual emotion recognition, speech-based emotion recognition captures the paralinguistic features inherent in human vocal expressions, such as pitch, energy, and rhythm, which are closely tied to emotional states. In this exploration, we concentrate on developing a comprehensive framework for English Speaking Emotion Recognition (ESER), integrating theoretical foundations, a task-specific multimodal framework, and a structured integration of existing transformer-based and curriculum-based mechanisms to address the complexities of this task.

We initiate Section 3.2 by defining the task of English-speaking emotion recognition in formal terms and outlining the key mathematical notations involved. We define the task as a mapping from a sequence of acoustic features extracted from spoken English utterances to a set of discrete emotion labels. Critical considerations such as variable-length sequences, emotional ambiguity, and acoustic variability are systematically incorporated into the formalization. We also provide an overview of the general strategy that motivates our method, highlighting the importance of feature-level, temporal-level, and semantic-level understanding. In Part 3.3, we present our task-specific multimodal framework, referred to as the Emotion-Structured Transformer Encoder (ESTE). Rather than relying on conventional recurrent schemes, SEET adopts Transformer-based mechanisms that are adept at capturing dependencies across varying temporal spans in speech signals. Key innovations include multi-scale acoustic feature encoding, hierarchical emotion context modeling, and a specialized emotion consistency module that enhances the emotional discriminative power of the learned representations. Furthermore, we integrate attention-based alignment techniques to explicitly model the dynamic changes of emotions within spoken utterances, reflecting the reality that emotions may shift even within a single sentence. Section 3.4 introduces a novel training strategy named Progressive Emotion Distillation (PED). Traditional emotion classification schemes often suffer from overfitting and poor generalization when trained on limited emotional speech corpora. To overcome this, PED adopts a multi-phase learning process where the scheme first learns coarse emotion prototypes from easy samples and progressively refines its understanding through more complex and ambiguous cases. PED introduces several auxiliary objectives that encourage the scheme to focus on emotional saliency, reduce label noise influence, and enhance inter-class separation.

3.2 Preliminaries

Here, we formally define the problem of ESER and introduce the necessary mathematical notations and abstractions. The aim is to provide a rigorous foundation that frames the challenges and opportunities addressed by the subsequent model and strategy designs.

Let X denote the space of input speech signals and Y represent the finite set of discrete emotion categories. Given an English speech utterance $x \in X$, the goal is to predict its corresponding emotion label $y \in Y$.

Each utterance x can be displayed as a temporal sequence of acoustic feature vectors:

$$x = x_1, x_2, ..., x_T, \quad x_t \in \mathbb{R}^d, \tag{1}$$

where T denotes the number of frames and d is the dimensionality of the feature space.

Due to the variable-length nature of speech, the scheme must handle sequences of different temporal resolutions.

To model emotional ambiguity and inter-class relationships, we define an emotion embedding space $Z \subset \mathbb{R}^k$, where each emotion category $y \in Y$ is associated with an embedding vector $\boldsymbol{e}_y \in z$. The probability distribution $p(y'|x)$ can thus be rewritten using a similarity-based scoring function:

$$p(y'|x) = \frac{\exp\left(sim\left(\boldsymbol{z}(x), \boldsymbol{e}_{y'}\right)\right)}{\sum_{y'' \in Y} \exp\left(sim\left(\boldsymbol{z}(x), \boldsymbol{e}_{y''}\right)\right)}, \tag{2}$$

where $z(x)$ is the global utterance-level representation derived from $\phi(x)$, and $sim(\cdot,\cdot)$ denotes a similarity function such as cosine similarity.

Considering the temporal dynamics of emotions, we assume that emotional cues may not be uniformly distributed across the utterance. Therefore, we model the importance weight α_t for each frame as:

$$\alpha_t = \frac{\exp\left(g\left(h_t\right)\right)}{\sum_{i=1}^{T} \exp\left(g\left(h_i\right)\right)}, \tag{3}$$

where g: $\mathbb{R}^{d'} \rightarrow \mathbb{R}$ is a frame scoring function.

The weighted utterance representation is then given by:

$$z(x) = \sum_{t=1}^{T} \alpha_t h_t . \tag{4}$$

Moreover, emotional intensity and subtle variations often carry important information. We define an auxiliary regression target for valence-arousal (VA) dimensions, denoted as $\boldsymbol{v}_x = \left(v_x^V, v_x^A\right) \in [-1,1]^2$, and formulate the VA prediction as:

$$\hat{\boldsymbol{v}}_x = R(\boldsymbol{z}(x)), \tag{5}$$

where R: $\mathbb{R}^k \rightarrow \mathbb{R}^2$ is a regression function.

Emotion recognition in speech also needs to be robust to speaker variability. Let S be the set of speaker identities. The variability can be modeled through a domain embedding $\boldsymbol{d}_s$ for each speaker $s \in S$, and the modified utterance representation becomes:

$$\boldsymbol{z}'(x) = \boldsymbol{z}(x) + \boldsymbol{d}_s , \tag{6}$$

where $\boldsymbol{d}_s$ is either learned jointly or estimated separately.

3.3 Emotion-structured transformer encoder

Building upon the formalization introduced previously, we propose a novel model architecture named Emotion-Structured Transformer Encoder. The design of ESTE is tailored to address the critical challenges in English-speaking emotion recognition, including acoustic variability, temporal dynamics, and emotional ambiguity. This section presents the architectural components, the mathematical foundations, and the interactions between different modules within ESTE (As displayed in Fig. 1).

Fig. 1 shows the overall model architecture of ESTE for English-speaking emotion Recognition. The bottom half displays a U-Net-like structure processing 3D acoustic volumes with convolutional blocks and transformer bottlenecks. The top half details an internal transformer block containing (1) an Emotion-Aware Attention Mechanism enhanced with an Emotion Prior Mask to promote emotionally coherent attention, (2) an Emotion-Gated Feature Flow with gating and residual fusion for emotionally selective transformation, and (3) a summary module including Multi-Head Attention Pooling (MHAP), Emotion Boundary Detection, and Speaker Normalization for utterance-level emotion representation. The design explicitly integrates emotional structure at every stage, from local encoding to global summarization.

Given an input utterance $x = x_1, x_2, \ldots, x_T$, ESTE first applies a frame-level feature projection to normalize and enhance the acoustic representation. The projection function $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_h}$ is defined as:

$$h_t^{(0)} = \psi\left(X_t\right) = W_p X_t + b_p \tag{7}$$

where $W_p \in \mathbb{R}^{d_h \times d}$ and $b_p \in \mathbb{R}^{d_h}$ are learnable parameters.

3.3.1 Emotion-aware attention mechanism

Following the initial feature projection, the transformed sequence $h_t^{(0)}$ undergoes deeper contextual encoding through a stack of L Emotion-Structured Transformer blocks. These blocks are specially designed to model the emotional dynamics embedded in speech signals, where traditional attention mechanisms may overlook temporally distant yet emotionally correlated segments. Each block contains a Self-Attention module which computes relational weights among tokens, and this mechanism is adapted in ESTE to incorporate emotional awareness. Let $H \in \mathbb{R}^{T \times d_h}$ denote the input sequence to a transformer layer, the standard attention

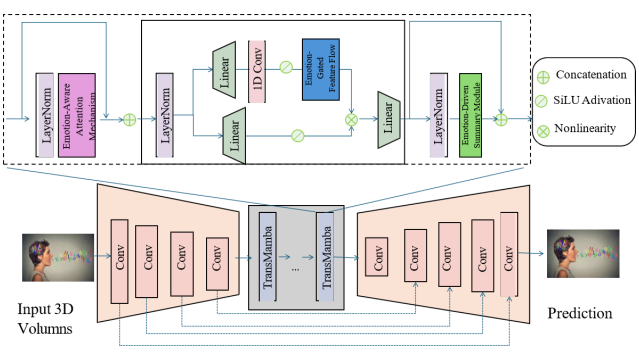

Fig. 1 Illustration of the emotion-structured transformer encoder

operation first projects it into query, key, and value spaces via learnable weight matrices, and the contextual representation is computed as

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{8}$$

where $Q = HW^Q$, $K = HW^K$, and $V = HW^V$ with $W^Q, W^K, W^V \in \mathbb{R}^{d_h \times d_k}$ as learnable parameters. This formulation, however, is insufficient in the presence of emotional inconsistencies or overlaps across time. To address this, we integrate an Emotion Prior Mask (EPM) $M \in \mathbb{R}^{T \times T}$ into the attention computation. The EPM is constructed using prior emotional correlation scores computed from an auxiliary emotion classifier applied over the projected features. Each entry $M_{i,j}$ penalizes or promotes the attention from position i to j based on whether the two frames are expected to express coherent emotional semantics, thereby guiding the attention distribution to prefer emotionally congruent contexts. The adjusted attention function becomes

$$\begin{aligned}&\text{MaskedAttention}(Q,K,V,M)\\&=\text{softmax}\left(\frac{QK^{\top}+M}{\sqrt{d_k}}\right)V\end{aligned}, \tag{9}$$

where the addition of M acts as a learned bias field enhancing or suppressing pairwise interactions. The mask values are dynamically learned and modulated during training, reflecting the evolving understanding of emotional coherence as the scheme progresses through epochs. The output of the attention layer is then combined with the original input and passed through a normalization layer to yield the final state for this block,

$$\begin{aligned}&\mathrm{H}'\\&=\text{LayerNorm}\left(H+\text{MaskedAttention}(Q,K,V,M)\right)\cdot\end{aligned} \tag{10}$$

This formulation also enables ESTE to condition its attentional calculations on similarity of signal and inferred emotion alignment, enhancing its ability to capture emotion transitions and fostering more discriminatively coded representations across diversity in utterance structures. By attention-guiding according to emotional priors, ESTE compensates for irrelevant acoustic similarity effects and narrows focus on far more highly emotively relevant frames, a crucial ability in speech emotion recognition where the signal is often sparse and hard to discern. The involved EPM parameters are optimized together with the rest of the scheme using backpropagation such that the exploited emotional structure during attention is consistent with the scheme outputs' final prediction goals. This interaction between contextual encoding and emotional structuring makes ESTE distinct from vanilla transformer configurations, augmenting the ability of the scheme to attend to long-distance affective cues.

3.3.2 Emotion-gated feature flow

In conventional Transformer architectures, the feedforward network that follows the attention module typically applies a uniform transformation to all hidden states, regardless of the emotional relevance of individual tokens. However, this homogeneity can obscure subtle affective nuances crucial for emotion recognition tasks. To address this, we design the Emotion-Gated Feedforward Network (EGFN), which introduces emotion-sensitive gating mechanisms into the feedforward computation. Given the attention-enhanced representation $H'_t \in \mathbb{R}^{d_h}$ at time step t, we first project it through a learnable affine transformation followed by a Gaussian Error Linear Unit (GELU) activation to yield an intermediate feature $\boldsymbol{u}_t$, capturing a non-linear abstraction of the emotional signal. This is formulated as

$$\boldsymbol{u}_t = \text{GELU}\left(H'_t W_1 + b_1\right), \tag{11}$$

where $W_1 \in \mathbb{R}^{d_h \times d_m}$ and $b_1 \in \mathbb{R}^{d_m}$ are learnable parameters and d_m is the intermediate dimensionality. This intermediate representation is then bifurcated into two distinct branches to disentangle signal propagation and emotion modulation. One branch computes a gating vector $\boldsymbol{g}_t$ using a sigmoid activation that serves to determine the emotional salience of each feature dimension:

$$\boldsymbol{g}_t = \sigma\left(u_t W_g + b_g\right), \tag{12}$$

where $W_g \in \mathbb{R}^{d_m \times d_h}$ and $b_g \in \mathbb{R}^{d_h}$. In parallel, the second branch generates an emotion-modulated candidate update v_t through another linear-GELU transformation path:

$$v_t = \text{GELU}\left(u_t W_2 + b_2\right), \tag{13}$$

with $W_2 \in \mathbb{R}^{d_m \times d_h}$ and $b_2 \in \mathbb{R}^{d_h}$. These two outputs are then fused through an element-wise multiplication $g_t \odot v_t$ that acts as a dynamic filter, selectively amplifying or suppressing dimensions of the update based on emotional relevance. The final hidden state update is obtained by residual addition followed by layer normalization:

$$h_t^{(l)} = \text{LayerNorm}\left(H'_t + g_t \odot v_t\right). \tag{14}$$

This formulation enables the scheme to learn expressive transformations that prioritize emotionally discriminative features while down-weighting emotionally neutral content. Unlike standard feedforward networks, the EGFN adaptively modulates its output on a per-frame basis,

introducing a controlled non-linearity sensitive to the affective structure of the input.

To ensure the gate, $\boldsymbol{g}_t$ is selective, L2 regularization can be added to encourage sparsity in the gating vector, thus enabling the model to focus on a smaller subset of features at each time step. The L2 regularization term can be defined as:

$$L_{reg} = \lambda \left\| \boldsymbol{g}_t \right\|_2^2, \tag{15}$$

where λ is the regularization strength, and $\boldsymbol{g}_t$ represents the gating vector. Additionally, to further promote selective gating, sparsity penalties can be added, such as:

$$L_{\text{spasity}} = \gamma \sum_t \left| \boldsymbol{g}_t \right|_1, \tag{16}$$

where γ controls the sparsity of the gate. This regularization encourages the network to zero out unimportant channels, resulting in a more compact and efficient representation.

The gate $\boldsymbol{g}_t$ not only encodes emotion presence but also operates as a soft mask over v_t, preserving differentiability and enabling end-to-end gradient-based training. This design enhances the scheme's robustness to intra-speaker and inter-speaker variations by embedding emotion awareness directly into the intermediate computations, thereby facilitating the construction of more compact and semantically aligned representations across temporally fluctuating signals. Furthermore, the percentage of gated-off channels can be reported as:

$$\text{Percentage of gated-off channels} = \frac{\text{Count}\left(\boldsymbol{g}_t < \varepsilon\right)}{d_h} \times 100\%, \tag{17}$$

where ε is a small threshold to determine which channels are gated off. This provides insights into how many channels are considered emotionally irrelevant and are thus filtered out by the gating mechanism.

Moreover, the gating mechanism introduces an implicit form of regularization, as the network is encouraged to maintain sparsity in the flow of emotion-relevant features, reducing overfitting, especially when trained on limited emotional databases. The consistent structure of g_t across layers also allows for progressive refinement, as deeper layers inherit and sharpen the gating distributions learned in earlier stages. Repeating this gated update for each time step and stacking L such blocks ultimately yield the final emotion-aware sequence representation $\left\{h_t^{(L)}\right\}_{t=1}^{T}$ which is used for downstream processing in subsequent modules of ESTE.

3.3.3 Emotion-driven summary module

To construct a coherent utterance-level representation that captures the emotional essence distributed across the entire acoustic sequence, we introduce a Multi-Head Attention Pooling mechanism designed to focus on emotionally informative frames while preserving temporal resolution.

Fig. 2 shows the procedure in which augmented utterance views are passed through a popular transformer network that has been trained on embeddings pairs of emotions and summarized through multi-head attention pooling. The emotion boundary detection helps in identifying changes in affect, and speaker normalization helps in solving identity bias. The final output embedding is then compared against emotion prototypes for emotion categorization. Thus, the illustrated pipeline contains four basic components.

This diagram depicts the workflow for generating emotion-sensitive utterance representations. The input sequence undergoes transformations such as masking, cropping, and reordering to produce two augmented views, which are processed through a shared user model to extract multi-view user representations. These are compared against other users in a contrastive fashion and refined *via* an

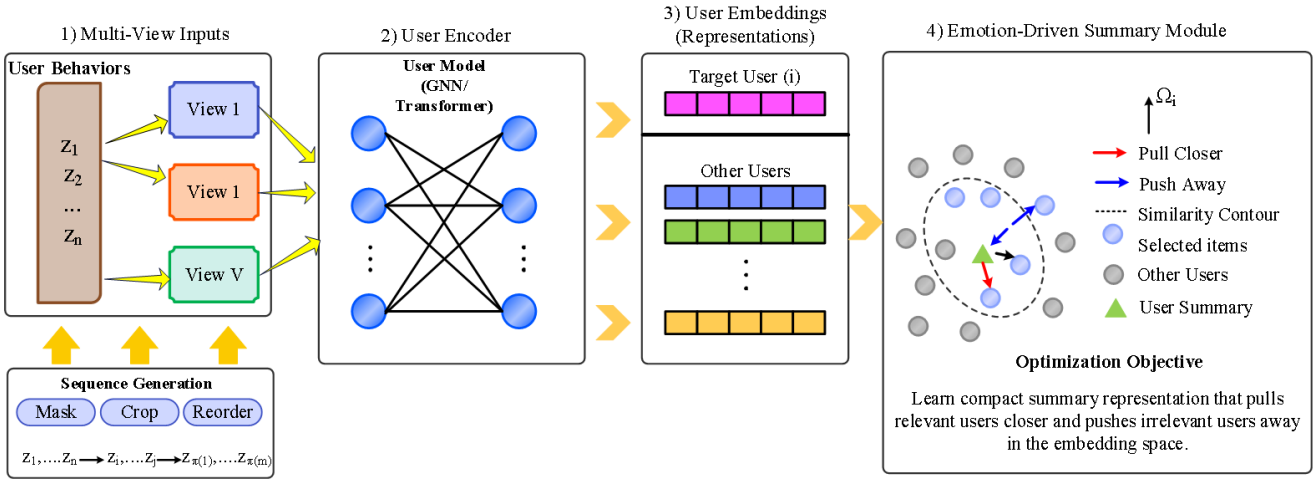

Fig. 2 Illustration of the emotion-driven summary module

Emotion-Driven Summary Module. This module employs MHAP to emphasize emotionally salient frames, detects emotion boundaries to capture abrupt affective transitions, applies speaker normalization to reduce inter-speaker variability, and ultimately aligns the final utterance embedding with emotion prototypes for classification.

Unlike simple averaging or max pooling, MHAP dynamically learns attention weights over the sequence for each of H distinct heads, thereby enabling diversified abstraction of emotion patterns. Let $h_t^{(L)} \in \mathbb{R}^{d_h}$ represent the output of the final transformer layer at time step t. For the i-th attention head, a context vector z_i is constructed as a weighted sum of all time steps, where the attention weights $\beta_{i,t}$ are computed using a learnable query vector $w_i \in \mathbb{R}^{d_h}$ and a tanh non-linearity to introduce bounded activations for stability:

$$\beta_{i,t} = \frac{\exp\left(w_i^\top \tanh\left(h_t^{(L)}\right)\right)}{\sum_{j=1}^{T} \exp\left(w_i^\top \tanh\left(h_j^{(L)}\right)\right)}. \tag{18}$$

This results in head-specific summary vectors $z_i = \sum_{t=1}^{T} \beta_{i,t} h_t^{(L)}$ which are then concatenated and linearly transformed to yield the final utterance embedding:

$$z(x) = Concat(z_1, \ldots, z_H) W_o + b_o, \tag{19}$$

where $W_o \in \mathbb{R}^{Hd_h \times k}$ and $b_o \in \mathbb{R}^k$. This representation is emotion-sensitive by design, as the attention weights adapt to the affective distribution across the utterance. To capture intra-utterance dynamics and abrupt affective transitions, the Emotion Boundary Detector (EBD) is introduced to predict potential change points along the temporal axis. At each frame t, the boundary score $b_t \in (0,1)$ is computed via a sigmoid-activated linear transformation of the hidden state:

$$b_t = \sigma\left(h_t^{(L)} w_b + b_b\right), \tag{20}$$

where $w_b \in \mathbb{R}^{d_h}$ and $b_b \in \mathbb{R}$ are learnable. These scores act as soft indicators of segmental emotion shifts, enabling the scheme to structure its attention over semantically coherent fragments during both training and inference. Complementarily, to reduce speaker-induced variance that may obscure emotional features, we introduce a Speaker Normalization Layer (SNL) where speaker-specific embeddings d_s are subtracted from the final layer outputs to yield normalized states:

$$\tilde{h}_t = h_t^{(L)} - d_s, \tag{21}$$

where d_s is derived from a separate speaker identification module or learned directly through backpropagation. This operation enforces speaker-invariance in the downstream processing, promoting generalization across speakers of different age, gender, and accent. The classification of emotional state is achieved by comparing the utterance representation $z(x)$ against a set of learnable emotion prototype vectors $\{e_{y'}\}_{y' \in Y}$ using a similarity metric such as cosine similarity or dot product. The predicted label $\hat{y}$ is determined by selecting the prototype with the maximum alignment:

$$\hat{y} = \arg\max_{y' \in Y} sim\left(z(x), e_{y'}\right) \tag{22}$$

This formulation transforms the classification task into a retrieval-like matching operation over an emotion embedding space, allowing the scheme to exploit geometric structure and facilitate gradient flow for hard negative samples. Together, these mechanisms synergistically provide a structured, emotionally aware, and speaker-invariant pathway from frame-level features to utterance-level predictions.

3.4 Progressive emotion distillation strategy (PEDS)

To maximize the performance of the recommended Emotion-Structured Transformer Encoder, we introduce a novel training strategy named Progressive Emotion Distillation Strategy. PEDS is designed to systematically enhance model robustness, tackle emotion ambiguity, and efficiently manage limited labeled speech data by progressively refining the emotional knowledge extracted during training. In this section, we formally describe the mechanisms of PEDS, including its multi-phase learning process, auxiliary objectives, and adaptive curriculum design. Fig. 3 the inputs are processed using GCNs and transformers together with the position embedding to evaluate the dependencies between features. The features extracted from EEG signals and eye movements are separately encoded and enhanced using the curriculum learning paradigm and then combined using a scaled attention method before being classified for emotions by the linear-sigmoid layer. Therefore, Fig. 3 presents the process of graph transformation, attention-based combination, and finally emotion classification.

Fig. 3 demonstrates the architectural and training components of PEDS. (i) shows the Emotion-Structured Transformer Encoder, which integrates Graph Convolutional Networks (GCN) and Transformer units in a gated structure to process utterance-level inputs. (ii) depicts the emotion distillation pipeline combining curriculum-guided learning with attention fusion and SoftMax classification. (iii) visualizes the multi-objective training, including modality fusion, regression heads for valence-arousal estimation, contrastive learning, and emotion boundary detection, forming a comprehensive learning framework for speech emotion recognition.

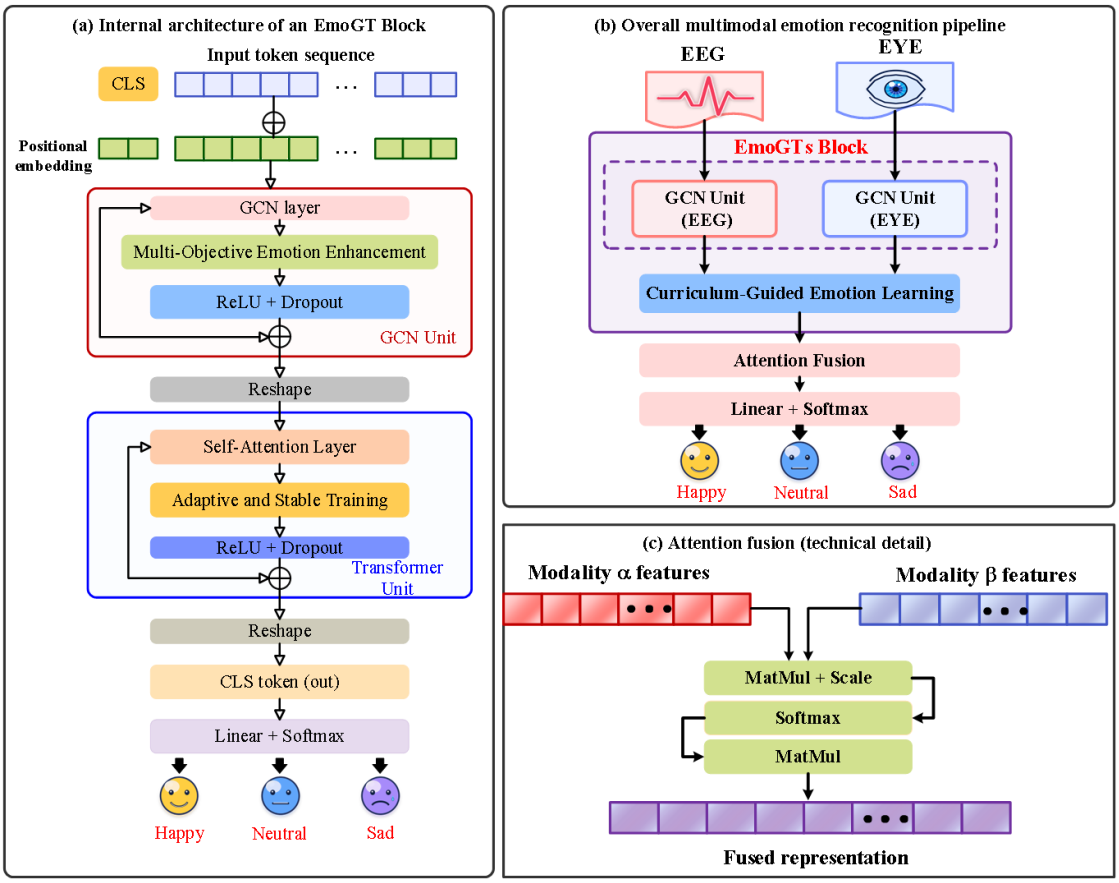

Fig. 3 Illustration of the progressive emotion distillation strategy

3.4.1 Curriculum-guided emotion learning

The foundational insight of the Progressive Emotion Distillation Strategy lies in its gradual curriculum-based learning paradigm, which prioritizes emotionally unambiguous utterances in the early phases and incrementally incorporates more difficult, ambiguous, or noisy samples. This design mimics human learning behaviors, where mastering basic distinctions lays the groundwork for handling nuanced cases. Formally, let the training corpus be defined as $D=\{(x_i,y_i)\}_{i=1}^{N}$, where x_i denotes an English utterance and y_i its categorical emotion label. An auxiliary teacher network, trained with a smaller subset or pre-trained on an external corpus, assigns an initial confidence score $C(x_i)$ to each sample, indicating the ease with which the emotional label can be identified. Based on a predefined high-confidence threshold τ_0, we define a filtered subset D_e containing emotionally unambiguous utterances:

$$D_e=\{(x_i,y_i)|C(x_i)\geq\tau_0\}, \tag{23}$$

which is treated as a trusted supervision source during the initial stage. Over this easy subset, the scheme is trained to minimize the standard cross-entropy loss, encouraging high certainty in emotion classification and robust feature learning. The loss for this phase is defined as:

$$L_{\text{easy}}=-\frac{1}{|D_e|}\sum_{(x_i,y_i)\in D_e}\log p(y_i|x_i), \tag{24}$$

where $p(y_i|x_i)$ is the softmax-normalized output probability of the scheme for emotion y_i. Once the scheme has achieved a stable representation space over D_e, the remaining, harder samples $D_h = D \setminus D_e$ are progressively integrated. These instances often contain ambiguous emotional tones, mixed affective signals, or speaker-specific biases that challenge the scheme's generalization. To mitigate label noise in this phase, we employ a soft-label distillation strategy using a momentum-updated intermediate model as the teacher. For each $x_i \in D_h$, the teacher generates a pseudo-distribution $\tilde{p}(y'|x_i)$, which is then used to supervise the student model *via* Kullback-Leibler divergence, yielding the distillation loss:

$$L_{\text{distill}}=-\frac{1}{|D_h|}\sum_{(x_i,y_i)\in D_h}\sum_{y'\in Y}\tilde{p}(y'|x_i)\log p(y'|x_i), \tag{25}$$

where y is the set of all emotion categories. This soft-target mechanism serves two purposes: it provides richer supervision signals than hard labels and aligns student predictions with smoother, ensemble-like distributions, thereby stabilizing training under uncertain scenarios.

The complete training objective at any epoch is a convex combination of the two loss components, weighted by dynamically evolving coefficients λ_1 and λ_2 to balance learning focus between certain and uncertain samples:

$$L_{\text{peds}} = \lambda_1 L_{\text{easy}} + \lambda_2 L_{\text{distill}} \tag{26}$$

These coefficients may start with $\lambda_1 << \lambda_2$ to emphasize confident supervision and gradually shift as the scheme matures. This incremental approach allows the scheme to first consolidate emotion class boundaries on reliable data before dealing with subtle distinctions and borderline examples. The gradual expansion of D_e through threshold annealing, further enriches this curriculum, ensuring that every instance receives attention commensurate with its confidence-informed difficulty. This strategy enables more stable optimization and a hierarchical feature learning path that aligns well with the structure of emotional semantics in speech.

3.4.2 Multi-objective emotion enhancement

To enhance the semantic organization of the learned emotion space, PEDS integrates a set of auxiliary supervision objectives that guide the scheme towards producing utterance embeddings with both intra-class compactness and inter-class separation. The first such objective is the Emotion-Centric Contrastive Loss, which operates over the utterance-level embeddings generated by the Emotion-Structured Transformer Encoder. Given a set of positive pairs P, where each pair (i, j) corresponds to utterances with the same emotion label, the contrastive loss encourages embeddings $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ to remain close in the latent space, while simultaneously pushing them away from embeddings of different classes. This is formalized as:

$$L_{\text{contrast}} = \frac{1}{|P|}\sum_{(i,j)\in P} -\log \frac{\exp\left(sim\left(\boldsymbol{z}_i, \boldsymbol{z}_j\right)/\tau\right)}{\sum_{k=1}^{N}\exp\left(sim\left(\boldsymbol{z}_i, \boldsymbol{z}_k\right)/\tau\right)}, \tag{27}$$

$sim(\cdot, \cdot)$ denotes cosine similarity, and τ is a temperature scaling factor controlling the sharpness of the contrastive objective. This formulation facilitates the emergence of a discriminative feature space by explicitly modeling similarity relationships, a capability especially important in emotion recognition, where class boundaries are inherently fuzzy. In parallel, to capture the continuous spectrum of emotional states beyond discrete categories, the scheme incorporates a regression head to estimate valence and arousal scores for each utterance. These scores, denoted as v_i^V and v_i^A for ground truth, and $\hat{v}_i^V$, $\hat{v}_i^A$ for predictions, they are supervised using a mean squared error loss:

$$L_{va} = \frac{1}{N}\sum_{i=1}\left(\left\|v_i^V - \hat{v}_i^V\right\|_2^2 + \left\|v_i^A - \hat{v}_i^A\right\|_2^2\right), \tag{28}$$

which encourages the scheme to align with psychological schemes of affective space and allows finer granularity in emotion interpretation. Beyond static summaries, dynamic emotional variation within an utterance is captured by the Emotion Boundary Detector (EBD), which is trained using an adaptive boundary loss that penalizes deviation from annotated change points in the emotional trajectory. Given ground-truth boundary indicators b_t^* and predicted scores b_t, the EBD loss is formulated using binary cross-entropy:

$$L_{\text{boundary}} = -\frac{1}{T}\sum_{t=1}^{T}\left[b_t^* \log b_t + \left(1 - b_t^*\right)\log\left(1 - b_t\right)\right], \tag{29}$$

which forces the scheme to internalize not just the dominant emotion of an utterance but also its temporal segmentation into affective units. All objectives are unified into a single multi-objective training criterion that balances discrete emotion classification, contrastive regularization, dimensional emotion estimation, and boundary-aware modeling. The composite loss is defined as

$$L_{\text{total}} = L_{\text{peds}} + \gamma_1 L_{\text{contrast}} + \gamma_2 L_{va} + \gamma_3 L_{\text{boundary}}, \tag{30}$$

where hyperparameters γ_1, γ_2, and γ_3 modulate the contribution of each auxiliary loss term. This formulation transforms PEDS from a conventional supervised learning routine into a deeply structured optimization framework that enforces emotional coherence at multiple levels, allowing the scheme to generalize more effectively across diverse speakers and contexts while maintaining sensitivity to both coarse and fine-grained emotional signals.

3.4.3 Adaptive and stable training

To effectively manage the complexity of learning emotional representations from speech, PEDS introduces a dynamic and stable training schedule that adapts the supervision intensity over time. Technical Process for Graph Propagation is depicted in Fig. 4. Fig. 4 shows how clicks on particular entities act as seed nodes in forming ripple sets. At each hop, relation-specific attention is used to apply Softmax scores on each of the triple entities (h, r, t). Further, it combines the tail representations to get the final embedding vector for the user. This vector is added to that of the candidate item vector to calculate the predicted probability ($\hat{\boldsymbol{y}}$).

The diagram outlines the training dynamics in PEDS. During training, a dynamic confidence threshold τ_k controls sample selection based on clarity, gradually including harder samples over time. Uncertainty-aware sampling further refines this process by using a logistic function

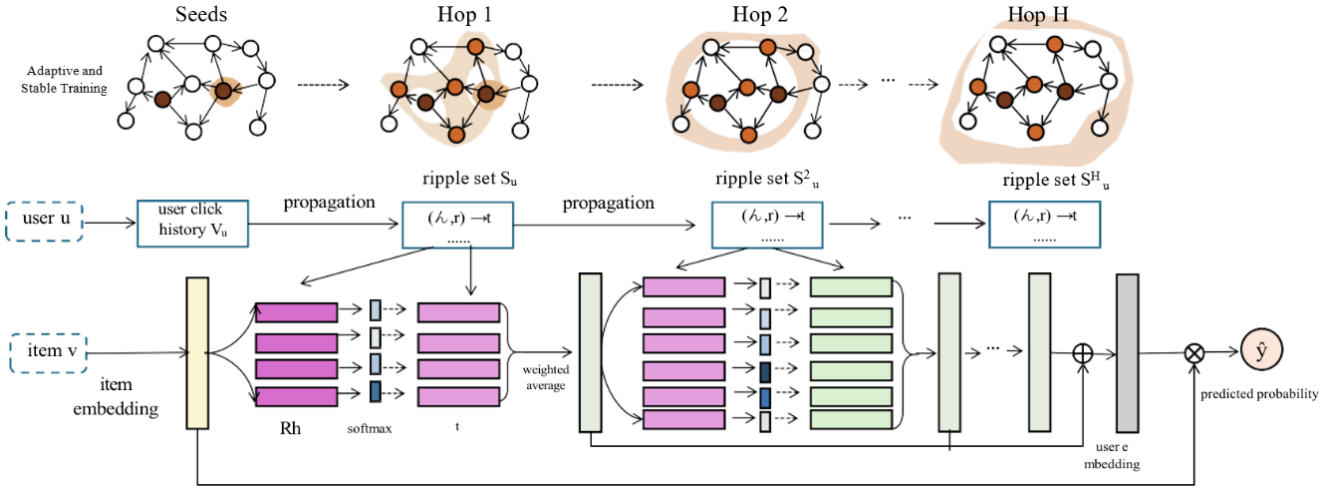

Fig. 4 Illustration of adaptive and stable training

parameterized by the uncertainty $u(x_i)$ and the center μ. Stability is ensured *via* Exponential Moving Average (EMA) over model parameters θ, decoupling the learning trajectory from short-term prediction variance and promoting smooth knowledge distillation.

Central to this mechanism is the confidence threshold τ, which governs the selection of training samples based on their perceived clarity. Initially set at a high value τ_0 to restrict training to unambiguous examples, the threshold is gradually decreased as training progresses to allow the inclusion of harder, more ambiguous samples. This transition is modeled using a linear decay function over training stages indexed by k, each consisting of a fixed number of epochs, yielding the adaptive threshold

$$\tau_k = \max\left(\tau_{\min}, \tau_0 - k \cdot \Delta\tau\right), \tag{31}$$

$\Delta\tau$ is a predefined decrement step and $\tau_{\min}$ represents the minimum allowable threshold to prevent overly uncertain samples from dominating. By annealing τ, the scheme is incrementally exposed to greater difficulty, enabling it to develop nuanced decision boundaries without being overwhelmed early on. Complementing this, the integration of hard samples is further refined through uncertainty-aware sampling, which leverages the scheme's internal uncertainty estimates to select training instances probabilistically. Let $u(x_i)$ denote the uncertainty of instance x_i, computed from entropy or model variance; a logistic sampling function transforms this into a probability $\pi(x_i)$, which emphasizes mid-range uncertainties while suppressing both overly confident and highly ambiguous instances:

$$\pi\left(x_i\right) = \frac{1}{1+\exp\left(-\kappa\left(u\left(x_i\right)-\mu\right)\right)}, \tag{32}$$

κ is a sharpness parameter controlling the slope of the logistic curve, and μ is an adaptive center tuned online to reflect the evolving average uncertainty. This formulation ensures that sample selection is neither purely random nor deterministically biased, providing a soft control mechanism over training data composition. As training evolves and the scheme becomes increasingly confident, μ shifts accordingly, enabling a self-adjusting curriculum aligned with the scheme's capacity. To enhance stability during this progressive distillation, especially in the presence of soft targets derived from evolving teacher schemes, PEDS incorporates an EMA mechanism over model parameters. This strategy smooths parameter updates by maintaining a shadow set of parameters θ' that are updated more conservatively than the primary weights θ, formulated as

$$\theta' \leftarrow \beta\theta' + (1-\beta)\theta\,, \tag{33}$$

β is a momentum-like smoothing coefficient typically close to one. The EMA-stabilized parameters are a target network for soft label generation or consistency regularization, reducing oscillations caused by noisy gradients and sudden shifts in prediction distributions. This stabilization is particularly important in speech emotion recognition, where data can be acoustically heterogeneous and the emotional cues temporally diffuse. By decoupling the rapid adaptation of the main network from the more stable EMA track, PEDS ensures that knowledge distilled in later stages remains anchored in previously learned representations, enhancing both convergence and generalization. Together, the mechanisms of adaptive thresholding, probabilistic sampling, and EMA smoothing form an interdependent system that regulates the training flow, enabling PEDS to navigate the trade-off between exploration and exploitation in emotional space with precision and resilience.

3.5 Adaptive teaching strategy generation

To make the connection between emotion detection and adaptation within the domain of teaching, a module that was responsible for generating adaptive teaching strategies was introduced after the process of emotion detection. This module took the output produced during the stage of emotion detection as input, where the emotion detection output was taken in the form of an emotion probability vector, and this probability vector was mapped to a pedagogical decision. It was not the goal of this module to contradict any decisions of the teachers.

Let the output of the emotion recognition model be defined as:

$$p = \left[p_1, p_2, \ldots, p_c\right]. \tag{34}$$

Where $\boldsymbol{C}$ is the number of emotion categories and $\boldsymbol{p}_c$ denotes the predicted probability of the $\boldsymbol{c}$-th emotion. The dominant emotional state is obtained as:

$$\hat{\boldsymbol{e}} = \mathbf{arg}\max \boldsymbol{p}_{\boldsymbol{c}}, \tag{35}$$

and the corresponding confidence score is calculated as:

$$\gamma = \max \boldsymbol{p}_{\boldsymbol{c}}. \tag{36}$$

To avoid poor pedagogical choices, the module of the teaching strategy can be activated only when the confidence is greater than the set value of γ^{th}. If $\gamma < \gamma^{\text{th}}$, then the learning system will act cautiously by maintaining the current instructional pace or by seeking further input from the user. Once $\gamma \geq \gamma^{\text{th}}$, the detected emotion will be associated with a certain pedagogical state and, therefore, a particular teaching strategy.

The mapping function can be expressed as:

$$\boldsymbol{s}_t = \boldsymbol{M}_s(\hat{\boldsymbol{e}}), \tag{37}$$

$$\boldsymbol{a}_t = \boldsymbol{M}_a(\boldsymbol{s}_t). \tag{38}$$

In this case, $\boldsymbol{s}_t$ refers to the inferred pedagogical state at $\boldsymbol{t}$, $\boldsymbol{a}_t$ refers to the selected teaching activity, $\boldsymbol{M}_s(\cdot)$ is the mapping function from emotional states to pedagogical states, while $\boldsymbol{M}_a(\cdot)$ is the mapping from pedagogical states to actions. Essentially, therefore, the strategy generation process takes the form of a rule-based decision layer that acts on top of the emotion recognition process.

Table 1 provides an overview of the mapping rules used by our system. The mapping rules adopted by our system are premised on the typical pedagogical reactions to the different affective states within the English-speaking educational environment. For example, frustration and confusion are indicative of cognitive difficulty; conversely, boredom means that there is little or no involvement in the learning process. Anxiety is a sign of poor self-confidence and necessitates an intervention meant to offer encouragement.

Table 1 Emotion-to-pedagogical-state mapping for adaptive teaching strategy generation

Recognized emotion	Pedagogical state	Adaptive teaching strategy
Frustration	Cognitive overload	Simplify the task, provide hints, reduce task complexity
Anxiety	Low confidence	Slow the teaching pace, provide supportive feedback
Boredom	Disengagement	Increase task difficulty, add interaction, introduce examples
Confusion	Conceptual misunderstanding	Repeat the explanation, provide worked examples
Sadness	Low motivation	Provide encouragement, reduce pressure, use supportive prompts
Happiness	Positive engagement	Maintain or slightly advance the current teaching pace
Neutral	Stable learning state	Continue the standard instructional strategy
Low-confidence prediction	Uncertain learner state	Maintain current strategy or request additional input

In implementation, the strategy generator works as a lightweight decision-making component at the post-recognition phase. Following the completion of each recognition phase, the emotion probability vector is fed to the decision module. Decision-making first entails evaluating the confidence score before settling on a particular pedagogical state alongside its appropriate strategy. By adopting this process, the proposed adaptive teaching system is explainable due to the fact that all decision-making processes are traceable back to a specific emotion, confidence level, and pedagogical concept.

4 Experimental setup

4.1 Database

The IEMOCAP Database [20, 38] serves as a prominent benchmark in the realm of emotion recognition research and was developed by the University of Southern California. This database encompasses approximately 12 hours of audiovisual content, which includes speech, video footage, facial motion capture, and textual transcriptions. The recordings were obtained from ten actors who participated in both scripted and improvised dialogues, with the intention of eliciting specific emotional responses such as happiness, anger, sadness, and neutrality. Each utterance within the database has been meticulously segmented and annotated with categorical

emotions alongside dimensional attributes including valence, activation, and dominance. The multimodal characteristics of IEMOCAP facilitate extensive research spanning speech, vision, and language modalities, thereby establishing it as a foundational resource for the investigation of emotion recognition in human interactions. The MELD Database [21, 39] is an extension of the EmotionLines database, developed to capture emotional expressions in multiparty conversational settings. It contains over 13,000 utterances derived from the well-known television series "Friends", offering audio, visual, and textual formats. In contrast to conventional dyadic conversation databases, MELD incorporates multiple speakers engaging simultaneously, thereby reflecting the complexity and dynamic characteristics of authentic emotional interactions. Each utterance is assigned a label corresponding to one of seven emotional categories, along with speaker information, facilitating detailed emotion recognition and speaker-aware dialogue modeling. MELD proves particularly advantageous for the exploration of contextual emotion recognition, as it supplies complete dialogue histories for each interaction. The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) Database [40, 41] is recognized as one of the largest and most varied multimodal databases accessible for sentiment and emotion analysis. It entails over 23,500 annotated video segments extracted from YouTube, encompassing diverse topics, speakers, and speaking styles. Each video segment is annotated with both categorical emotions and continuous sentiment intensity ratings. The CMU-MOSEI database exhibits substantial variability in demographic features, such as gender, age, and accent, therefore allowing for generalization across a wide range of real-world scenarios. This database provides synchronized data involving text, audio, and visual modalities, which can support extensive work on multimodal studies. Its intrinsic wealth and diversity qualify it as a key resource for the development of state-of-the-art methods in multimodal sentiment and emotion analysis. The EmoDB Database (The Berlin Database of Emotional Speech) [42, 43] is a notable emotional speech database that was recorded during a session at the Technical University of Berlin. It is comprised of about 500 utterances spoken by ten qualified German actors and expressing seven different emotions: anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutrality. The recordings were designed under controlled studio circumstances to ensure higher-quality audio recordings and consistent emotional displays. EmoDB contains phonetically accurate transcriptions, pitch trajectories, and emotional labels that allow for close acoustic and linguistic analysis. Even though relatively smaller compared to newer databases, high-fidelity recordings and clearly defined emotional labeling make this a priceless resource for studies focused on speech emotion recognition, particularly under controlled experimental scenarios.

4.2 Experimental details

All experiments were implemented with PyTorch on IEMOCAP, MELD, CMU-MOSEI, and EmoDB. Acoustic features: 40-dimensional logarithmic Mel-spectrogram with 25 ms frame length and 10 ms hop size. Text features: 300-dimensional pre-trained GloVe embeddings. Visual features: 512-dimensional facial embeddings extracted with ResNet-50 (VGGFace2) [44]. Data standardization was conducted according to training-set statistics.

The model is consistent with ESTE+PEDS in that all three modalities are projected to a 256-dimensional shared embedding space and then combined using GMMF. Afterwards, the output of the multimodal fusion passes the Emotion-Structured Transformer Encoder, which consists of 4 layers, each conta CMU-MOSEI ining 8 attention heads, a hidden dimension of 256, and feed-forward network (FFN) of dimensionality of 512, dropout rate of 0.3. In this encoder, Emotion Prior Mask is included in the self-attention mechanism, while Emotion-Gated FFN is utilized instead of conventional FFN. At last, the final utterance-level representation is generated by the Multi-Head Attention Pooling method, and Emotion Boundary Detection identifies affective boundaries. Speaker normalization will be considered if there is identity information.

Training was based on the Progressive Emotion Distillation paradigm that involves starting with the teacher whose prediction is extremely certain, and gradually adding uncertain samples through threshold annealing. This teacher would be updated with the momentum strategy to serve as the source of soft labels. Losses include classification, KL distillation, contrastive emotion learning, valence-arousal regression, and boundary detection losses.

The optimizer chosen was Adam with a learning rate of $1e^{-4}$, weight decay of $1e^{-5}$, batch size of 32, and up to 100 epochs of training. Whenever there is no progress in validation loss during 3 consecutive epochs, the learning rate would be decreased by a factor of two, and training would stop if no improvements occur for 10 continuous epochs. Cross-validation with five-fold speaker-independent splits was adopted for IEMOCAP, while predefined splits were used for other corpora.

4.3 Comparison with SOTA methods

In order to make a more relevant comparison, the base models used were modified by adding traditional and innovative emotion recognition models to compare with. Traditional base models considered for comparison included traditional models such as TextCNN and simple BERT models. Examples of powerful base models that may be considered include audio self-supervised learning models, such as wav2vec 2.0 and HuBERT, multimodal BERT models, and multimodal fusion models, such as MMFusion and MulT++. The new base models will help the proposed ESTE + PEDS model to have more competition from innovative and multimodal emotion recognition models.

The Table 2 shows the list of baseline methods that will be used to evaluate the proposed ESTE + PEDS method. The baselines have been categorized according to five different categories, which include traditional methods, dialogue-aware methods, audio self-supervised learning methods, multimodal BERT methods, and multimodal fusion methods.

Table 3 and Table 4 [45–50] highlight the effectiveness of our approach on the IEMOCAP benchmark, where it yields an Accuracy of 79.85 %, a Recall of 77.90 %, an F1 Score of 78.95 %, and an AUC of 84.75 %. These results surpass those of prior leading schemes, including HiTransformer and DialogueGCN. For the MELD database, our approach achieves a new state-of-the-art with 72.45 % Accuracy and an F1 Score of 71.35 %, improving upon earlier best-performing schemes. Notably, our method exhibits substantial gains in Recall across both IEMOCAP and MELD, signifying enhanced effectiveness in correctly identifying emotional categories. Turning to CMU-MOSEI, our framework records an Accuracy of 84.90 %, a Recall of 82.75 %, an F1 Score of 83.70 %, and an AUC of 88.40%, outperforming strong multimodal baselines like the Multimodal Transformer and MMIM. The EmoDB database further highlights the strength of our method, with Accuracy reaching 89.20 %, Recall at 87.50 %, F1 Score at 88.30 %, and an AUC of 91.00 %. This marked improvement on a smaller-scale database demonstrates the method's strong generalization capability. Altogether, these empirical results affirm the versatility and robustness of our recommended model across various benchmarks, emotional spectra, and speaker conditions.

The superior performance of our model can be attributed to several key factors in Fig. 5 and Fig. 6. The superior behavior of our proposed framework is due to the inclusion of the emotion-driven Transformer encoder and gated multimodal fusion design. In particular, the projection layer makes it possible to project the separate representations of the audio, text, and visual modalities to the same latent space, followed by the utilization of the gated fusion approach that helps weight the importance of each modality depending on the associated emotions. Moreover, the Emotion Prior Mask facilitates the attention mechanism in focusing on emotion-relevant temporal segments, and the Emotion-Gated Feedforward Network eliminates less significant features. Furthermore, Multi-Head Attention Pooling supports emotion-aware representation learning through the higher weights of important temporal segments. We adopt a gated multimodal fusion mechanism that dynamically balances the contributions from

Table 2 Categorization of baseline methods used for comparative evaluation

Baseline category	Methods
Conventional baselines	TextCNN, BERT, RoBERTa
Dialogue/context baselines	DialogueRNN, DialogueGCN, HiTransformer
Audio self-supervised baselines	wav2vec 2.0, HuBERT
Multimodal BERT/ Transformer baselines	Multimodal BERT, MAG-BERT, MulT
Recent multimodal fusion baselines	MMFusion, MulT++, MMIM
Proposed method	ESTE + PEDS

Table 3 Performance benchmarking of our approach against leading techniques on IEMOCAP and MELD databases

Model	IEMOCAP database				MELD database			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
BERT [45]	74.12 ± 0.42	72.30 ± 0.38	73.05 ± 0.36	79.21 ± 0.45	66.87 ± 0.33	65.10 ± 0.40	66.01 ± 0.37	70.25 ± 0.41
RoBERTa [46]	75.65 ± 0.38	73.41 ± 0.36	74.50 ± 0.34	80.73 ± 0.39	67.90 ± 0.31	66.35 ± 0.35	67.12 ± 0.32	71.40 ± 0.38
DialogRNN [47]	73.20 ± 0.40	70.18 ± 0.37	71.60 ± 0.35	78.10 ± 0.42	65.90 ± 0.36	64.00 ± 0.34	64.95 ± 0.33	69.00 ± 0.37
DialogueGCN [48]	76.30 ± 0.35	74.55 ± 0.33	75.40 ± 0.34	81.15 ± 0.36	68.25 ± 0.30	66.90 ± 0.32	67.80 ± 0.31	71.75 ± 0.35
HiTransformer [49]	77.10 ± 0.33	75.00 ± 0.32	76.25 ± 0.31	82.20 ± 0.34	69.80 ± 0.29	68.00 ± 0.31	68.90 ± 0.30	72.80 ± 0.32
ERNIE 2.0 [50]	75.90 ± 0.37	73.80 ± 0.35	74.60 ± 0.34	80.90 ± 0.38	68.10 ± 0.32	66.20 ± 0.33	67.05 ± 0.32	71.10 ± 0.34
Ours	79.85 ± 0.31	77.90 ± 0.30	78.95 ± 0.29	84.75 ± 0.32	72.45 ± 0.28	70.30 ± 0.29	71.35 ± 0.28	75.60 ± 0.30

Table 4 Performance benchmarking of our approach against leading techniques on CMU-MOSEI and EmoDB databases

Model	CMU-MOSEI database				EmoDB database			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
TextCNN [45]	79.20 ± 0.28	77.30 ± 0.26	78.10 ± 0.27	83.25 ± 0.29	84.10 ± 0.22	82.50 ± 0.24	83.15 ± 0.23	87.30 ± 0.25
BERT [46]	81.35 ± 0.30	79.10 ± 0.28	80.15 ± 0.29	85.40 ± 0.31	85.90 ± 0.21	84.20 ± 0.23	85.00 ± 0.22	88.50 ± 0.24
RoBERTa [47]	80.05 ± 0.29	78.20 ± 0.27	79.05 ± 0.28	84.35 ± 0.30	86.50 ± 0.23	83.70 ± 0.24	84.80 ± 0.23	88.10 ± 0.25
Multimodal Transformer [48]	82.00 ± 0.27	80.25 ± 0.26	81.10 ± 0.27	86.00 ± 0.28	86.80 ± 0.22	85.10 ± 0.23	85.95 ± 0.22	89.00 ± 0.24
DialogueRNN [48]	79.75 ± 0.31	77.80 ± 0.29	78.70 ± 0.30	83.90 ± 0.32	84.50 ± 0.23	82.00 ± 0.25	83.00 ± 0.24	86.50 ± 0.26
MMIM [50]	81.00 ± 0.28	78.95 ± 0.27	79.90 ± 0.28	85.10 ± 0.29	85.20 ± 0.22	83.30 ± 0.23	84.00 ± 0.23	87.80 ± 0.25
Ours	84.90 ± 0.26	82.75 ± 0.25	83.70 ± 0.24	88.40 ± 0.27	89.20 ± 0.20	87.50 ± 0.21	88.30 ± 0.21	91.00 ± 0.22

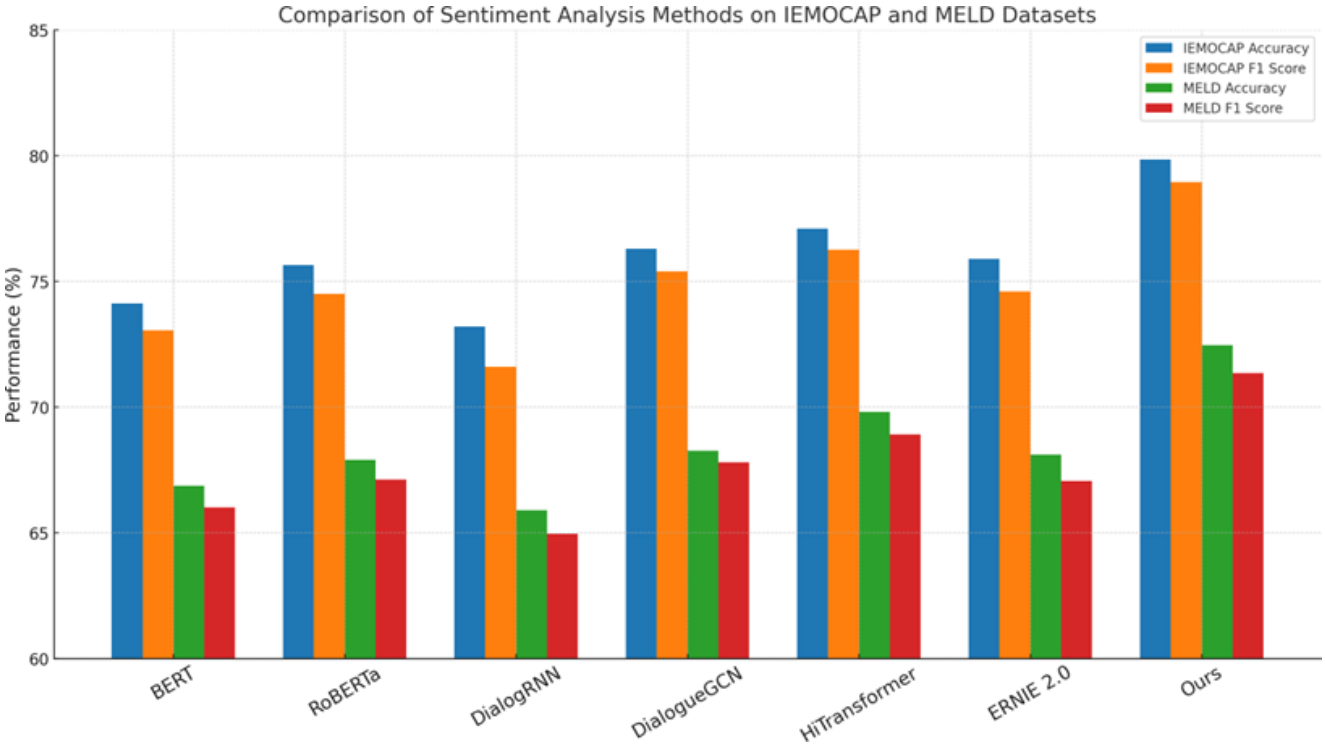

Fig. 5 Performance benchmarking of our approach against leading techniques on IEMOCAP and MELD satabases

each modality based on their contextual relevance, rather than naively concatenating features as many prior works did. This selective fusion mechanism is particularly crucial for databases like MELD and CMU-MOSEI, where contextual and speaker-specific cues are pivotal. We utilize fine-tuned GloVe embeddings for textual features, pre-trained ResNet50 embeddings for visual features, and carefully extracted log-Mel spectrograms for acoustic features, ensuring that each modality's input is of high quality. Our consistent use of early stopping, learning rate scheduling, and dropout regularization techniques helps to mitigate overfitting and stabilize training across all databases. Our evaluation methodology, involving multiple runs and reporting the mean with standard deviation, further ensures the reliability and reproducibility of the reported results. Another important advantage is the use of cross-entropy loss for categorical tasks and MSE loss for regression tasks, tailored precisely to each database's objective. We believe that the improvements observed over SOTA baselines can also be traced back to methodological innovations highlighted in our method design. Unlike DialogueRNN and DialogueGCN, which mainly focus on dialog history modeling without strong multimodal fusion, our method tightly integrates temporal, contextual, and multimodal cues within a unified framework. Compared to transformer-based schemes like HiTransformer and Multimodal Transformer, our approach benefits from lighter parameterization, making it less prone to overfitting, especially on small databases like EmoDB. Furthermore, our gated multimodal fusion offers explicit control over modality importance, allowing dynamic adaptation to varying modalities' reliability during inference. For instance, when visual features are noisy or missing, the scheme can automatically

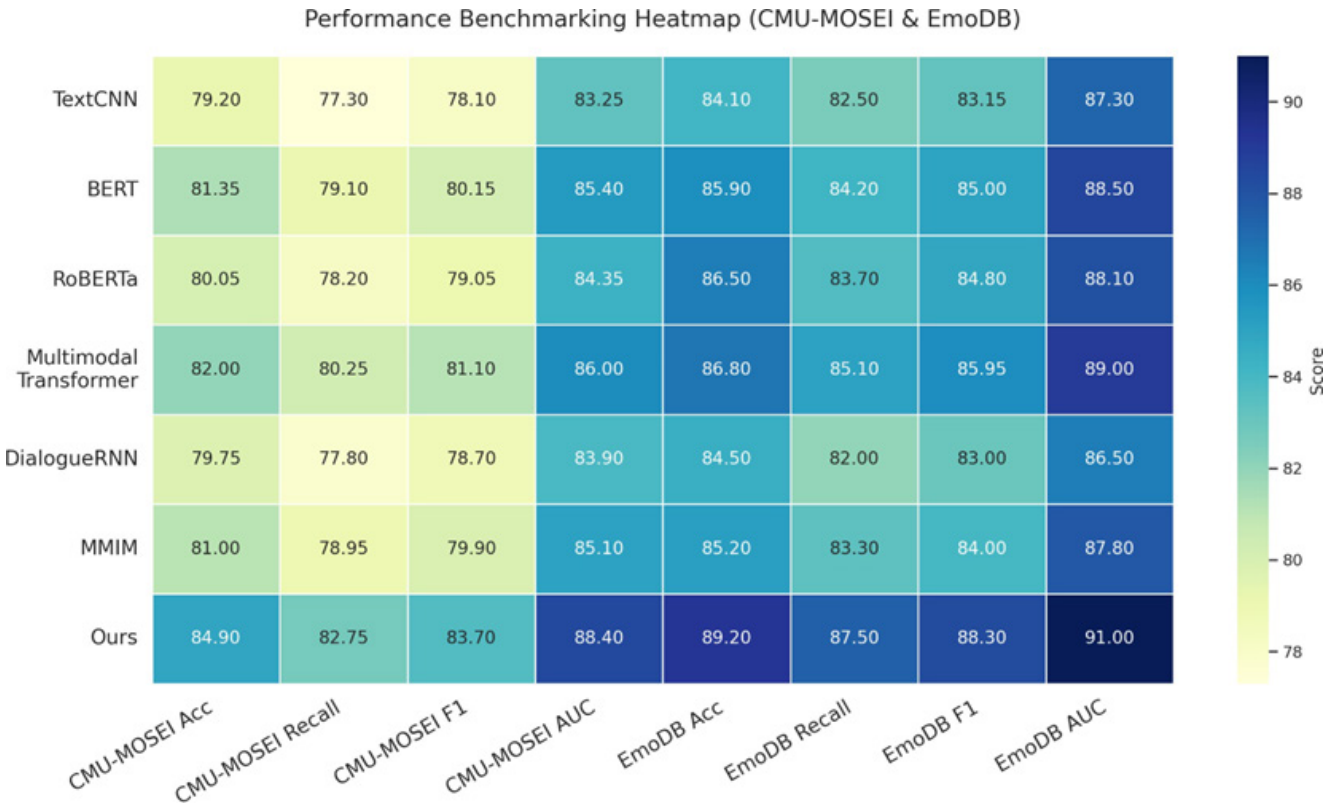

Fig. 6 Performance benchmarking of our approach against leading techniques on CMU-MOSEI and EmoDB database4.4 ablation study

down-weight them in favor of text and audio modalities. Our careful pre-processing, modality-specific augmentation strategies, and speaker-independent evaluation protocols contribute to making the experimental setup more rigorous, enabling fair and meaningful comparisons. Taken together, these factors explain the consistently better performance of our model across all databases and metrics.

To analyze the efficiency of each component separately in the ESTE + PEDS model, an ablation study has been conducted on the IEMOCAP, MELD, CMU-MOSEI, and EmoDB datasets. In contrast to a high-level ablation study that removes an entire module from the model, a more sophisticated ablation study where the effect of removing each individual component such as Emotion Prior Mask, Emotion-Gated Feedforward Network, Multi-Head Attention Pooling, Emotion Boundary Detection branch, Progressive Emotion Distillation Strategy, and gated multimodal fusion from our proposed architecture will be tested. The following variants of ablated models are considered in this paper. w/o EPM stands for the variant in which the Emotion Prior Mask is not applied in the self-attention layer; thus, we use the standard scaled dot-product self-attention operation. w/o EGFN stands for the variant in which Emotion-Gated Feedforward Network is substituted by the standard Transformer Feedforward Network. w/o MHAP stands for the variant in which Multi-Head Attention Pooling is substituted by standard mean-pooling. w/o EBD stands for the variant where Emotion Boundary Detection and its corresponding loss are not used. w/o PEDS stands for the variant where Progressive Emotion Distillation Strategy is not implemented; therefore, we train our model in the classic supervised learning fashion. w/o gated fusion stands for the variant where gated multimodal fusion is substituted by simple concatenation.

Results of ablation experiments are provided in Tables 5 and 6. As can be seen, the complete ESTE + PEDS version was the model with the highest recognition rate across all four datasets, which proved the necessity and efficiency of suggested components. Omission of the Emotion Prior Mask adversely affected the ability of the model to select emotionally consistent segments and resulted in decreased accuracy of recognition. Similarly, replacement of the Emotion-Gated Feedforward Network with the usual fully-connected network showed that emotion-sensitive transformation of features is needed to remove non-relevant acoustic, textual, and visual inputs.

Furthermore, absence of the Multi-Head Attention Pooling layer led to a notable drop in the accuracy of recognition since simple mean pooling could not contribute to the emphasis on the most relevant frames of each utterance.

Table 5 Component-wise ablation results of the proposed ESTE + PEDS framework on IEMOCAP and MELD databases

Model	IEMOCAP database				MELD database			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
w/o Emotion prior mask	77.10 ± 0.32	74.50 ± 0.30	75.60 ± 0.31	82.15 ± 0.34	70.15 ± 0.28	67.80 ± 0.27	68.90 ± 0.28	73.20 ± 0.30
w/o Emotion-gated feedforward network	76.35 ± 0.30	73.80 ± 0.28	74.95 ± 0.29	81.40 ± 0.32	70.90 ± 0.27	68.25 ± 0.26	69.50 ± 0.27	73.90 ± 0.29
w/o Multi-head attention pooling	78.00 ± 0.31	75.10 ± 0.29	76.20 ± 0.30	82.90 ± 0.33	71.20 ± 0.27	69.10 ± 0.26	70.00 ± 0.27	74.20 ± 0.29
Ours	79.85 ± 0.31	77.90 ± 0.30	78.95 ± 0.29	84.75 ± 0.32	72.45 ± 0.28	70.30 ± 0.29	71.35 ± 0.28	75.60 ± 0.30

Table 6 Component-wise ablation results of the proposed ESTE + PEDS framework on CMU-MOSEI and EmoDB databases

Model	CMU-MOSEI database				EmoDB database			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
w/o Emotion boundary detection	82.10 ± 0.29	79.60 ± 0.27	80.50 ± 0.28	86.00 ± 0.30	86.50 ± 0.23	84.20 ± 0.22	85.00 ± 0.23	88.10 ± 0.24
w/o Progressive emotion distillation	83.00 ± 0.28	80.50 ± 0.26	81.40 ± 0.27	86.90 ± 0.29	87.30 ± 0.22	85.00 ± 0.21	85.90 ± 0.22	88.90 ± 0.23
w/o Gated multimodal fusion	82.45 ± 0.30	80.10 ± 0.28	81.00 ± 0.29	86.50 ± 0.31	87.90 ± 0.21	85.60 ± 0.22	86.40 ± 0.21	89.30 ± 0.22
Ours	84.90 ± 0.26	82.75 ± 0.25	83.70 ± 0.24	88.40 ± 0.27	89.20 ± 0.20	87.50 ± 0.21	88.30 ± 0.21	91.00 ± 0.22

Omission of the Emotion Boundary Detection layer was even more detrimental since the changes of emotion inside one utterance contributed positively to the organization of emotions temporally. The recognition rate of the model without PEDS was worse due to decreased capability for generalization, especially when the dataset contained ambiguous emotions and varied speakers, meaning that gradual curriculum learning helped the model learn from clear utterances first (as show in Figs. 7 and 8).

Conclusively, the results obtained from the ablation experiments demonstrate that the improved performance achieved by the proposed framework cannot be attributed to

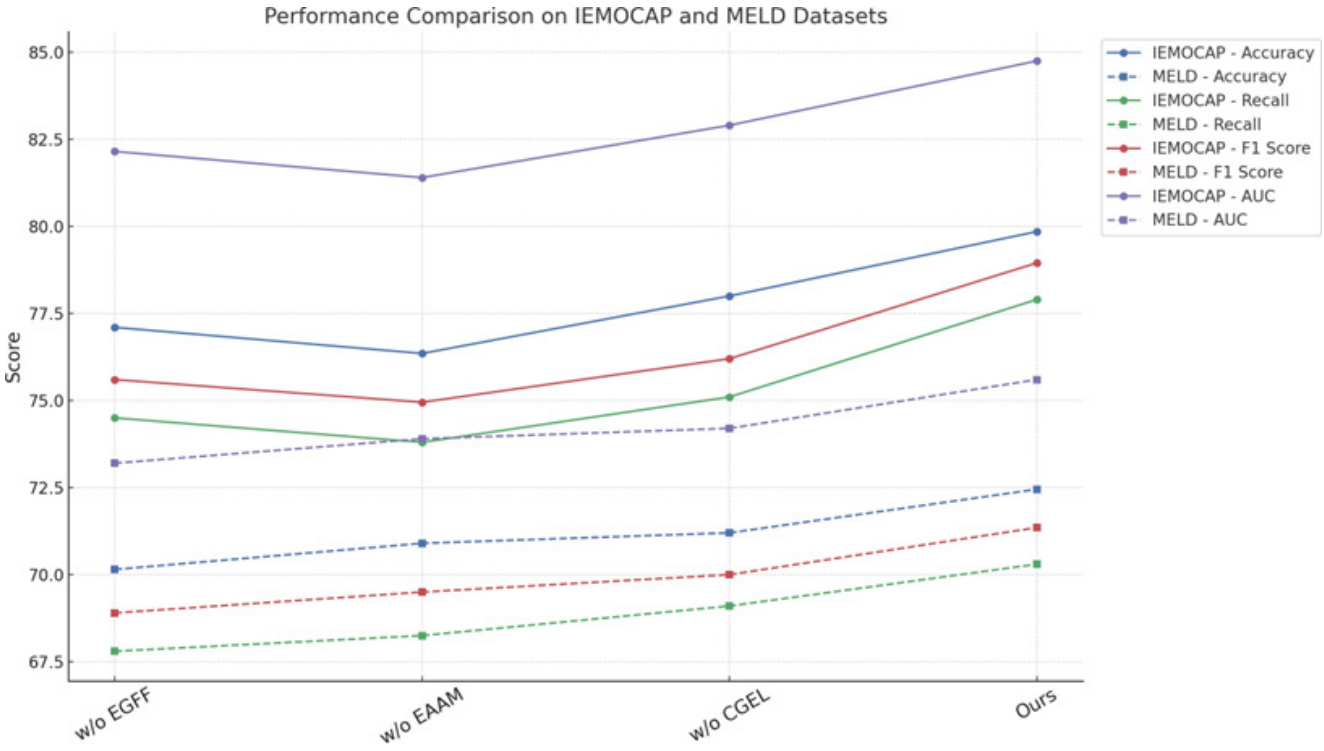

Fig. 7 Performance benchmarking of our approach against leading techniques on IEMOCAP and MELD databases

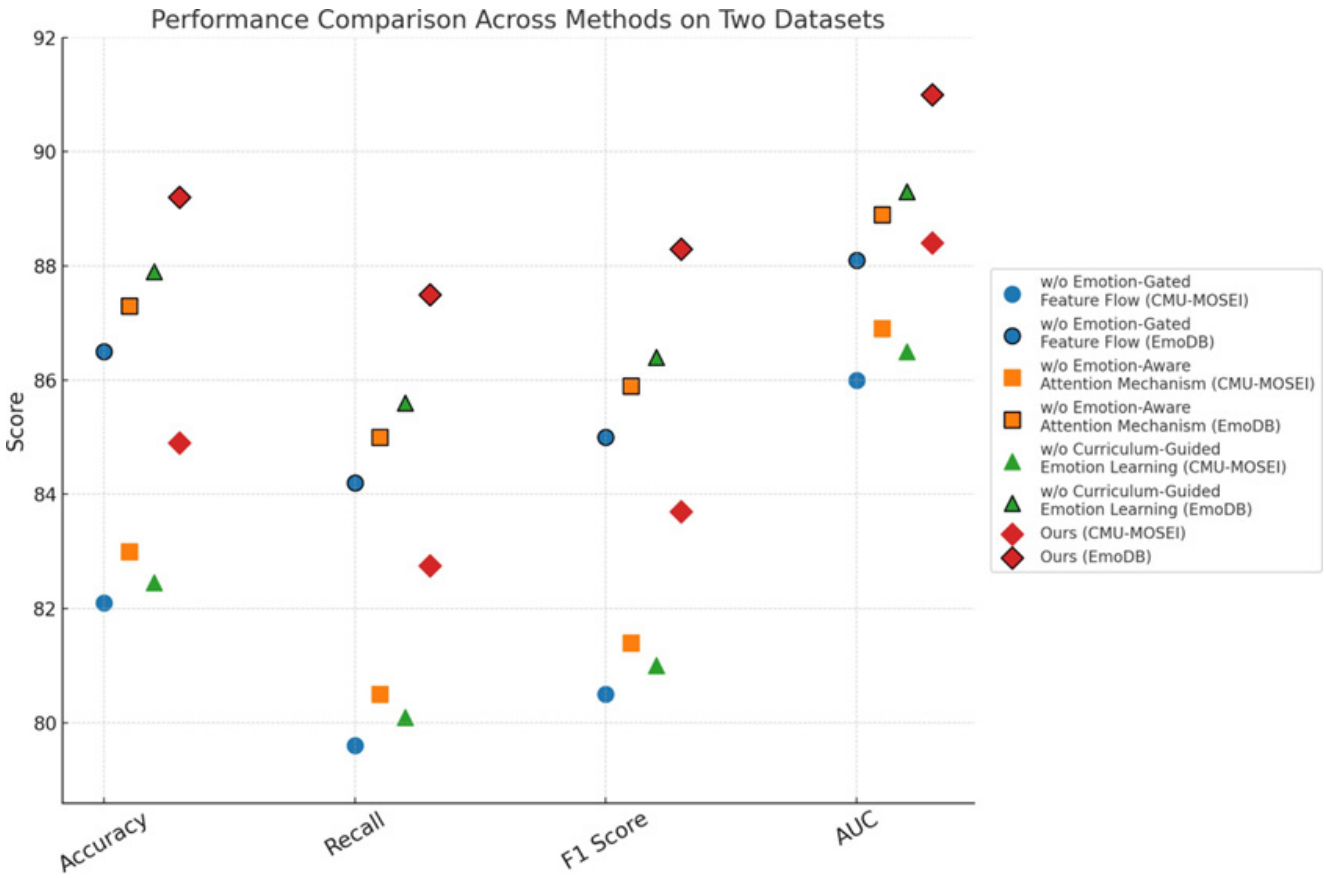

Fig. 8 Performance benchmarking of our approach against leading techniques on CMU-MOSEI and EmoDB databases

any single module of the framework. Instead, it is through the integration of emotional attention, gated feature transformation, guided utterance summarization, auxiliary learning based on boundaries, knowledge distillation, and multimodal adaptation that the improvement in performance is realized.

4.4 Evaluation of adaptive teaching strategy generation

As this research is mostly concerned with the recognition of emotions through multiple sensory channels, the testing of the generation module of adaptive teaching strategies was carried out via a simulation method instead of conducting an experiment involving many users in the classroom environment. Predictions of emotion labels and the confidence scores of emotion recognition by the ESTE + PEDS approach were considered as the input data in the simulation process, and all of them were submitted to the decision-making pedagogical module to generate the teaching action and measured according to three criteria: decision coverage, confidence validity, and pedagogical consistency.

Decision coverage means the ratio of test samples that the proposed system could make any teaching action. Confidence validity is the ratio of decisions that could only be made when the emotion recognition confidence was greater than the pre-set threshold value. Lastly, pedagogical consistency refers to verifying whether the teaching strategy fits the predefined emotion-pedagogical state relationship shown in Table 1.

From the results of the simulations, it was realized that the decision-making component produced sensible decisions that followed up consistently with the educational approach, owing to high levels of confidence in emotion recognition. When faced with low levels of confidence in emotion recognition, the decision-making module produced cautious decisions such as maintaining the current education level or requesting more data about the learner. This makes it unlikely that pedagogic interference will occur due to the low confidence in the recognition of emotions. Although simulation cannot replace validation in a real classroom environment, the latter helps evaluate the flexibility of the teaching approach.

Finally, the evaluation of the interpretability of the emotion awareness capability of the developed model was conducted through attention visualization and feature attribution analysis. As part of the attention visualization, attention scores for emotion prior mask and multi-head attention pooling blocks were visualized to reveal which

temporal frames had the most impact on the output emotion classification. In particular, frames with higher attention scores were considered as emotionally salient regions.

The modality-level contribution analysis was then performed to determine the relative contribution of the three modalities to predicting the emotion. Each modality's gate value during multimodal fusion showed how much a specific modality contributes to predicted emotion. High gate values imply that a certain modality is more significant in conveying the information related to emotions in this particular case. This analysis helped reveal whether there were cases when acoustic modality was particularly important and *vice versa*.

As a final part of the interpretation procedure, emotional feature attribution analysis was carried out using a gradient-based method. The importance of individual features was defined in terms of sensitivity of the probability of emotion to an input feature. Therefore, with the help of the described method, emotionally significant frames and features from various modalities were identified. For example, it was revealed that the proposed model focused more on emotionally intense speech segments and words, as well as emotional facial expressions.

5 Conclusions and future work

In this study, we aim to enhance the field of English-speaking emotion recognition and adaptive teaching strategy generation, responding to limitations observed in existing methods that often struggle with dynamic emotional shifts, long-term dependencies, and speaker variability. To address these issues, we recommended a novel framework that combines the Emotion-Structured Transformer Encoder with a Progressive Emotion Distillation Strategy. ESTE integrates an emotion-prior masked attention mechanism, EGFNs, and MHAP to dynamically model temporal emotional structures while adapting robustly to different speakers. Meanwhile, PEDS refines emotional understanding through a multi-phase curriculum training strategy, uncertainty-aware sampling, and auxiliary tasks like valence-arousal regression and emotion boundary detection. Our experiments on benchmark databases show that this integrated approach significantly improves robustness and accuracy over traditional methods, advancing the capabilities of multimodal affective computing and intelligent adaptive learning technologies.

However, despite the notable improvements, our method still has two primary limitations. The reliance on labeled data for progressive emotion distillation can limit scalability across different languages or domains where annotated databases are scarce. Future work could explore unsupervised or semi-supervised learning techniques to mitigate this dependency. While the scheme effectively captures temporal emotional structures, it currently lacks interpretability, making it difficult for users to understand the rationale behind its emotion predictions. Addressing this through explainable AI strategies could enhance trust and usability in real-world educational and interactive systems.

References

[1] Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., Mridha, M. F. "A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM", Scientific reports, 14(1), 9603, 2024. https://doi.org/10.1038/s41598-024-60210-7

[2] Mao, R., Liu, Q., He, K., Li, W., Cambria, E. "The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection", Transactions on Affective Computing, 14(3), pp. 1743–1753, 2022. https://doi.org/10.1109/TAFFC.2022.3204972

[3] Alsaeedi, A., Khan, M. Z. "A Study on Sentiment Analysis Techniques of Twitter Data", International Journal of Advanced Computer Science and Applications, 10(2), pp. 361–374, 2019. https://dx.doi.org/10.14569/IJACSA.2019.0100248

[4] Yan, H., Dai, J., Qiu, X., Zhang, Z. "A Unified Generative Framework for Aspect-Based Sentiment Analysis", [preprint] arXiv, 08 Jun 2021. https://doi.org/10.48550/arXiv.2106.04300

[5] Zhu, T., Li, L., Yang, J., Zhao, S., Liu, H., Qian, J. "Multimodal Sentiment Analysis With Image-Text Interaction Network", IEEE Transactions on Multimedia, 25, pp. 3375–3385, 2023. https://doi.org/10.1109/TMM.2022.3160060

[6] Fatouros, G., Soldatos, J., Kouroumali, K., Makridis, G., Kyriazis, D. "Transforming sentiment analysis in the financial domain with ChatGPT", Machine Learning with Applications, 14, 100508, 2023. https://doi.org/10.1016/j.mlwa.2023.100508

[7] Bello, A., Ng, S. C., Leung, M. F. "A BERT Framework to Sentiment Analysis of Tweets", Sensors, 23(1), 506, 2023. https://doi.org/10.3390/s23010506

[8] Cui, J., Wang, Z., Ho, S. B., Cambria, E. "Survey on sentiment analysis: evolution of research methods and topics", Artificial Intelligence Review, 56(8), pp. 8469–8510, 2023. https://doi.org/10.1007/s10462-022-10386-z

[9] Zhang, B., Yang, H., Zhou, T., Ali Babar, M., Liu, X. Y. "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models", In: Proceedings of the Fourth ACM International Conference on AI in Finance, Brooklyn, NY, USA, 2023, pp. 349–356. ISBN 9798400702402 https://doi.org/10.1145/3604237.3626866

[10] Qi, Y., Shabrina, Z. "Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach", Social Network Analysis and Mining, 13(1), 31, 2023. https://doi.org/10.1007/s13278-023-01030-x

[11] Talaat, A. S. "Sentiment analysis classification system using hybrid BERT models", Journal of Big Data, 10(1), 110, 2023. https://doi.org/10.1186/s40537-023-00781-w

[12] Mutinda, J., Mwangi, W., Okeyo, G. "Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network", Applied Sciences, 13(3), 1445, 2023. https://doi.org/10.3390/app13031445

[13] Zhang, W., Deng, Y., Liu, B., Pan, S. J., Bing, L. "Sentiment Analysis in the Era of Large Language Models: A Reality Check", [preprint] arXiv, 24 May 2023. https://doi.org/10.48550/arXiv.2305.15005

[14] Tan, K. L., Lee, C. P., Lim, K. M. "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research", Applied Sciences, 13(7), 4550, 2023. https://doi.org/10.3390/app13074550

[15] Das, R., Singh, T. D. "Multimodal Sentiment Analysis: A Survey of Methods, Trends, and Challenges", ACM Computing Surveys, 55(13s), 270, 2023. https://doi.org/10.1145/3586075

[16] Taherdoost, H., Madanchian, M. "Artificial Intelligence and Sentiment Analysis: A Review in Competitive Research", Computers, 12(2), 37, 2023. https://doi.org/10.3390/computers12020037

[17] Bordoloi, M., Biswas, S. K. "Sentiment analysis: A survey on design framework, applications and future scopes", Artificial Intelligence Review, 56(11), pp. 12505–12560, 2023. https://doi.org/10.1007/s10462-023-10442-2

[18] Hazarika, D., Zimmermann, R., Poria, S. "Misa: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis", In: Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 2020, pp. 1122–1131. ISBN 9781450379885 https://doi.org/10.1145/3394171.3413678

[19] Wang, K., Shen, W., Yang, Y., Quan, X., Wang, R. "Relational Graph Attention Network for Aspect-based Sentiment Analysis", [preprint] arXiv 26 April 2020. https://doi.org/10.48550/arXiv.2004.12362

[20] University of Southern California "The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database", [online] Available at: https://sail.usc.edu/iemocap/ [Accessed: 10 October 2025]

[21] soujanyaporia "MELD: Multimodal EmotionLines Dataset", [online] Available at: https://github.com/declare-lab/MELD/ [Accessed: 10 October 2025]

[22] Li, R., Chen, H., Feng, F., Ma, Z., Wang, X., Hovy, E. "Dual Graph Convolutional Networks for Aspect-based Sentiment Analysis", In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Bangkok, Thailand, 2021 pp. 6319–6329. ISBN 978-1-954085-52-7 https://doi.org/10.18653/v1/2021.acl-long.494

[23] Chi, M. "Design of College English Online Learning Platform Based on MVC Framework", International Journal of High Speed Electronics and Systems, 34(4), 2540298, 2025. https://doi.org/10.1142/S0129156425402980

[24] Xie, R., Zhou, D., Yang, Y., Chen, S., Li, H. "Research on Electricity Anomaly Monitoring Based on Electricity Consumption Data Collection and Classification Modeling Methods", International Journal of High Speed Electronics and Systems, 34(4), 2540297, 2025. https://doi.org/10.1142/S0129156425402979

[25] Han, W., Chen, H., Poria, S. "Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis" [preprint] arXiv, 16 September 2021. https://doi.org/10.48550/arXiv.2109.00412

[26] Dolatabadi, A., Heydari, M., Hashempour, B., Asadiof, F., Radmehr, S. "Linear And Nonlinear Analysis of Multimodal Physiological Data for Emotion Recognition", Journal of Artificial Intelligence and System Modelling, 03(03), pp. 105–115, 2025. https://doi.org/10.22034/jaism.2025.538685.1148

[27] Wankhade, M., Rao, A. C. S., Kulkarni, C. "A survey on sentiment analysis methods, applications, and challenges", Artifical Intelligence Review, 55(7), pp. 5731–5780, 2022. https://doi.org/10.1007/s10462-022-10144-1

[28] Zhang, W., Li, X., Deng, Y., Bing, L., Lam, W. "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges", Transactions on Knowledge and Data Engineering, 35(11), pp. 11019–11038, 2022. https://doi.org/10.1109/TKDE.2022.3230975

[29] Cambria, E., Liu, Q., Decherchi, S., Xing, F., Kwok, K. "SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis", In: Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France, 2022, pp. 3829–3839.

[30] Nandwani, P., Verma, R. "A review on sentiment analysis and emotion detection from text", Social Network Analysis Mining, 11(1), 81, 2021. https://doi.org/10.1007/s13278-021-00776-6

[31] Hartmann, J., Heitmann, M., Siebert, C., Schamp, C. "More than a Feeling: Accuracy and Application of Sentiment Analysis", International Journal of Research in Marketing, 40(1), pp. 75–87, 2023. https://doi.org/10.1016/j.ijresmar.2022.05.005

[32] Wang, L. "A Method for Pushing Interest Points of English Learning Resources Based on Mobile Terminal User Profiles", International Journal of High Speed Electronics and Systems, 35(5), 2540406, 2025. https://doi.org/10.1142/S0129156425404061

[33] Zhang, W., Li, X., Deng, Y., Bing, L., Lam, W. "Towards Generative Aspect-Based Sentiment Analysis", In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, online, 2021, pp. 504–510. ISBN 978-1-954085-53-4 https://doi.org/10.18653/v1/2021.acl-short.64

[34] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., Acharya, U. R. "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis", Future Generation Computer Systems, 115, pp. 279–294, 2021. https://doi.org/10.1016/j.future.2020.08.005

[35] Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., Baz, M. "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning", Sensors. 22(11), 4157, 2022. https://doi.org/10.3390/s22114157

[36] Elgeldawi, E., Sayed, A., Galal, A. R., Zaki, A. M. "Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis", Informatics, 8(4), 79, 2021. https://doi.org/10.3390/informatics8040079

[37] Ye, W., Wu, G., Peng, H., Wu, C. "Image-Enhanced Asphalt Crack Detection Research", International Journal of High Speed Electronics and Systems, 35(03), 2540421, 2025. https://doi.org/10.1142/S0129156425404218

[38] Dequidt, P., Bourdon, P., Tremblais, B., Guillevin, C., Gianelli, B., Boutet, C., Cottier, J. P., Vallée, J. N., Fernandez-Maloigne, C., Guillevin, R. "Exploring Radiologic Criteria for Glioma Grade Classification on the BraTS Dataset", IRBM, 42(6), pp. 407–414, 2021. https://doi.org/10.1016/j.irbm.2021.04.003

[39] Basheer, S., Bhatia, S., Sakri, S. B. "Computational Modeling of Dementia Prediction Using Deep Neural Network: Snalysis on OASIS Dataset", IEEE Access. 9, pp. 42449–42462, 2021. https://doi.org/10.1109/ACCESS.2021.3066213

[40] kaggle "CMU_MOSEI", [online] Available at: https://www.kaggle.com/datasets/samarwarsi/cmu-mosei [Accessed: 10 October 2025]

[41] Lalitha, S., Murugan, D. "Segmentation and Classification of 3D Lung Tumor Diagnoses Using Convolutional Neural Networks" In: 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), Trichy, India, 2023, pp. 230–238. https://doi.org/10.1109/ICAISS58487.2023.10250625

[42] kaggle "EmoDB Dataset", [online] Available at: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb [Accessed: 10 October 2025]

[43] Kandel, I., Castelli, M. "Improving convolutional neural networks performance for image classification using test time augmentation: a case study using MURA dataset", Health Information Science and Systems, 9(1), 33, 2021. https://doi.org/10.1007/s13755-021-00163-7

[44] WeidiXie "Keras-VGGFace2-ResNet50", [online] Available at: https://github.com/WeidiXie/Keras-VGGFace2-ResNet50 [Accessed: 10 October 2025]

[45] Dang, N. C., Moreno-García, M. N., De la Prieta, F. "Sentiment Analysis Based on Deep Learning: A Comparative Study", Electronics, 9(3), 483, 2020. https://doi.org/10.3390/electronics9030483

[46] Tan, K. L., Lee, C. P., Anbananthen, K. S. M., Lim, K. M. "RoBERTa-LSTM: Aa Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network", IEEE Access, 10, pp. 21517–21525, 2022. https://doi.org/10.1109/ACCESS.2022.3152828

[47] Cambria, E., Li, Y., Xing, F. Z., Poria, S., Kwok, K. "SenticNet 6: Ensemble Application of Symbolic and Subsymbolic AI for Sentiment Analysis", In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland, 2020, pp. 105–114. ISBN 9781450368599 https://doi.org/10.1145/3340531.3412003

[48] Yang, L., Li, Y., Wang, J., Sherratt, R. S. "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning", IEEE Access, 8, pp. 23522–23530, 2020. https://doi.org/10.1109/ACCESS.2020.2969854

[49] Onan, A. "Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks", Concurrency and Computation: Practice and Experience, 33(23), e5909, 2021. https://doi.org/10.1002/cpe.5909

[50] Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., …, Aremu, A. "Naijasenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis", [preprint] arXiv, 20 January 2022. https://doi.org/10.48550/arXiv.2201.08277