

# Spatiotemporal Self-Attentive Graph-TCN Framework for Enhanced Teen Stress Detection

Indumathy P.<sup>1\*</sup>, Praveen Kumar R.<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Easwari Engineering College, Ramapuram, Chennai, Tamil Nadu 600089, India

<sup>2</sup>Electronics and Communication Engineering, Easwari Engineering College, Ramapuram, Chennai, Tamil Nadu 600089, India

\*Email: indumathyp56@gmail.com

Received for publication  
August 19, 2025

## Abstract

Timely and precise detection of adolescent stress is critical for early intervention, yet current CNN- or Transformer-based models fail to jointly capture both structural facial dependencies and temporal emotion dynamics. To address this gap, we propose a novel spatiotemporal self-attentive graph-TCN (STG-TCN) framework that integrates graph-based spatial reasoning with temporal convolutional modeling. Unlike conventional approaches that concatenate spatial and temporal features, our method employs a cross-attention fusion mechanism to align and enrich these modalities. The proposed model was evaluated on the dataset for affective states in e-environments (DAiSEE) under strict cross-subject protocols, achieving 90.2% classification accuracy, a 12% improvement over graph attention network (GAT)-only baselines, and a 9% improvement over temporal convolutional networks (TCN)-only baselines. Furthermore, STG-TCN reduced false negatives by 20%, demonstrating high sensitivity to subtle stress cues such as micro-expressions and transient muscle contractions. These findings establish STG-TCN as the first framework specifically tailored for adolescent stress recognition in real-world video settings, thereby advancing the field of affective computing with both methodological innovation and practical implications for unobtrusive mental health monitoring.

## Keywords

stress detection, adolescents, graph attention network, temporal convolutional network, EfficientNet-Lite, facial landmarks, spatiotemporal fusion, deep learning, cross-attention

## 1. Introduction

Mental pressure in adolescence is becoming a severe public health problem that, if undiagnosed, can have psychological as well as physiological consequences for life [1]. The dynamic and sometimes cryptic nature of stress-related behaviors in teenagers makes the automated recognition problematic [2]. Traditional methods, which are mainly based on static CNNs or sequence architectures, may achieve suboptimal results, because they have a certain limitation in learning spatial facial structures as well as the evolution of emotional cues along the temporal dimension. It demands a framework that fully represents the

structural dependencies among facial parts and, at the same time, captures temporal emotion dynamics with great accuracy [3].

Adolescents will show more low-level, situational stress indicators, sometimes mixed with confusion, frustration, or withdrawal. Such affective states are often expressed as micro-expressions, muscle constriction, and temporal facial activity patterns [4]. The available techniques only separate spatial or temporal rather than synthesizing the interactions that are important for identifying some of the subtle or ambivalent stress signs. A more efficient computational model should not only segment and weight facial regions in a smart manner but should also be able to

learn the context-dependent temporal transitions that are characteristic of adolescent stressed expressions in different emotional states [5].

Graph-based techniques are also popular for modeling structured data, such as for facial landmark analysis, where there are meaningful spatial relationships between keypoints [6]. However, the history of the neighbors is not captured explicitly, and sometimes significant relational features could be diluted in a traditional convolution due to the process of uniform aggregation of the neighborhood. This limitation can be addressed by attention mechanisms on graphs, which allow for context-aware node interaction weighting. Meanwhile, temporal convolutional networks (TCNs) are appealing RNN alternatives that can model hierarchical temporal dependencies without the gradient vanishing problem, and are especially appealing for long-term facial behavior modeling [7]. Nevertheless, only a small number of studies have successfully integrated the two perspectives in a single framework for early adolescent stress detection.

Such a multifaceted challenge is tackled by the combination of graph attention and temporal convolution modeling. The framework first extracts per-frame embeddings using a lightweight facial encoder and then projects the embeddings onto a facial landmark graph for spatial operation using an attention-based model [8]. In parallel, a temporal stream captures the evolution of the frame-wise embedding. The two types of modalities are synchronously integrated by a cross-attention fusion, which correlates structural cues with act sequences and yields a strong spatiotemporal representation. This hybrid approach enables a more sensitive and interpretable stress detection in real-world adolescent video data, even in cases of low-intensity or mixed-affective states.

## a. Background

Facial analysis technologies have progressed significantly with the integration of convolutional and attention-based networks, primarily focusing on adult emotion recognition. However, adolescent faces exhibit more ambiguous expressions and often lack the pronounced features typical in adult datasets, reducing the effectiveness of standard models. Stress, unlike emotion, manifests subtly and sporadically, making it harder to detect without a model attuned to both spatial structure and time-series behavior.

Previous datasets and models in stress recognition have either focused on physiological signals or on general affective states without distinguishing age-based expression differences. In visual-only

settings, models often misclassify stress as confusion or fatigue due to overlapping facial signals. This research fills a critical gap by building a model tailored for adolescent facial video streams, focusing specifically on stress-related cues validated through affective proxies.

## b. Scope and motivation

This study centers on facial-video-based stress detection in adolescents using the dataset for affective states in e-environments (DAiSEE). It is limited to visual data without multimodal inputs and emphasizes spatiotemporal learning via deep neural architectures. The approach targets generalization across subjects and sessions through structural and temporal alignment techniques. Adolescents under stress often go undiagnosed due to the fleeting and nuanced nature of their facial expressions. Early detection through passive video analysis can support intervention without requiring active cooperation. This work aims to design a scalable and accurate system that can recognize stress based on naturalistic facial behavior.

## c. Objectives and key contributions

The primary objective is to develop a unified deep learning framework that jointly models spatial and temporal aspects of adolescent facial expressions linked to stress. This involves integrating graph-based attention mechanisms for localized spatial reasoning with temporal convolution for sequential pattern analysis. The model is rigorously evaluated under cross-subject protocols on a real-world dataset to ensure generalization and robustness. The key contributions of the paper are as follows:

- Introduces a Graph Attention + Temporal Convolution framework, specifically for adolescent stress recognition using facial videos;
- Develops a cross-attention fusion module that aligns graph-based and sequence-based features to enhance prediction accuracy;
- Validates on the DAiSEE dataset, achieving 90.2% accuracy under cross-subject settings, outperforming conventional spatial or temporal baselines.

While our earlier works have explored facial analysis using deep learning architectures, this study is distinct in both scope and methodological design. Previous approaches were limited to either spatial or temporal feature modeling, often with simple concatenation of modalities. By contrast, the present

work introduces a cross-attention fusion mechanism within the Spatio-Temporal Graph-based Temporal Convolutional Network (STG-TCN) framework, which uniquely aligns graph-based landmark reasoning with temporal sequence modeling. This design not only enhances interpretability but also addresses the specific challenge of detecting stress in adolescents, a demographic largely overlooked in prior publications. Thus, the novelty of this study lies in its integration of spatial-temporal attention and its application to adolescent stress detection under cross-subject settings, which differentiates it from our earlier contributions.

#### d. Organization of the paper

The paper begins with a review of the related works, highlighting advancements in facial stress detection, graph-based affective modeling, and temporal learning approaches. Then, the methodology section details the proposed STG-TCN framework, explaining its embedding strategy, graph attention mechanism, temporal convolutional components, and fusion design. Next, the experimentation section presents the dataset characteristics, cross-subject evaluation protocol, and training configuration. This is followed by results and discussion, offering both quantitative performance metrics and interpretative analysis through ablation studies and baseline comparisons. Finally, the conclusion summarizes the key findings and proposes future directions, including multimodal integration and domain adaptation enhancements.

## II. Related Works

Seo et al. [9] proposed a deep learning-based stress detection model using multimodal inputs, including ECG, respiration, and facial features, to assess work-related stress. Their approach demonstrated reasonable accuracy, particularly when fusing respiration data with facial landmark coordinates. However, performance declined in multilevel stress classification due to overlapping class distributions, as revealed through *t*-SNE analysis. A key limitation lies in the reliance on physiological sensors, which may hinder scalability and unobtrusive real-world deployment. Campanella et al. [10] introduced a stress detection method using signals from the Empatica E4 wearable device, analyzed through machine learning models like Random Forest, SVM, and logistic regression. Their approach enabled real-time monitoring via physiological features, achieving the highest accuracy of 76.5% with Random Forest. The key advantage lies in the system's portability and

continuous tracking capability using non-invasive sensors. Yet, the generalizability across larger populations may be limited because the method is based on handcrafted features and small training samples.

Li and Lima [11] implemented a facial emotion recognition technique adopting ResNet-50 to extract deep features, with the objective of enhancing the performance of such a model in human-computer interface scenarios. They improved accuracy compared with traditional CNN models by using residual learning for advanced feature representation. One benefit of this approach is its enhanced detection performance on benchmark datasets. But it still has many limitations in robustness and generalization for different kinds of facial expressions, especially for in-the-wild facial expressions. Gupta et al. [12] designed a real-time facial emotional deep learning-informed system for checking online learners' engaged reactions. The method employs models such as Inception-V3, VGG19, and ResNet-50 to recognize emotions and to calculate an engagement index. ResNet-50 is the best-performing model, with a 92.3% accuracy, and is well-suited for real-time applications. Yet focusing on facial expression alone would ignore the influence of situational and behavioral elements on engagement, rendering it less comprehensive in the wide variety of learning contexts.

Mohan et al. [13] presented FER-net, a convnet modeling task to recognize facial expression over handcrafted feature-based approaches. The model was compared with 21 state-of-the-art methods on five standard datasets and obtained competitive performance, especially on complex emotion categories. Its robustness is that it works well with deep learning to capture subtle facial expression variations without the necessity of handcrafted features. But differences in performance among datasets imply that there may be limitations to generalization to very diverse or non-task-constrained real-world scenes. Khattak et al. [14] proposed a deep learning-based method for facial emotion recognition, attempting to compensate for the drawbacks from incorrect selection of the CNN layers in the prior works. The CNN model they offer is able to effectively classify emotion, age, and gender from facial expressions with high accuracy. Another of the major advantages is the strong results obtained for several recognition tasks: 95.65% for emotions, 98.5% for age, and 99.14% for gender. Nevertheless, the model in the method greatly depends on static facial images and thus cannot well describe the temporal changing information of facial expression, which is a foundation for dynamic affective analysis in real video environments.

Savchenko et al. [15] presented a video-based facial analysis pipeline to analyze student interaction in e-learning. The approach integrates face detection/tracking and a trained neural network for real-time prediction of individual emotions, engagement, and group affect. A major strength is its ability to run efficiently on mobile devices without cloud processing, preserving privacy. However, reliance on pretrained static image models may limit its sensitivity to dynamic emotional shifts common in continuous video streams. Minaee et al. [16] proposed Deep-Emotion, a facial expression recognition model built on attentional convolutional networks to address high intraclass variation in emotional cues. By focusing on the most informative facial regions, the model outperformed traditional handcrafted feature-based approaches across several challenging datasets like FER-2013 and JAFFE. A key advantage is its use of attention mechanisms that enhance interpretability and robustness to partial occlusions. However, its dependency on labeled datasets and the absence of temporal modeling limit its applicability for real-time dynamic emotion recognition scenarios.

Furthermore, recent works have extended stress recognition to diverse contexts. For example, Tuan et al. [33] analyzed depression, anxiety, and stress detection among recovered coronavirus

disease (COVID-19) patients using machine learning approaches, while another study by Tuan et al. [34] explored stress expressions in online communities via subreddit analysis. These studies underline the growing interest in stress-related computational modeling and provide additional support for addressing stress in adolescent populations through advanced deep learning frameworks.

### a. Research gap

Despite growing interest in automated stress detection, existing models struggle to generalize across real-world, dynamic scenarios involving adolescents. Many rely on physiological sensors or static image-based emotion recognition, which either hinder scalability or fail to capture temporal emotional cues. Models using handcrafted features often lack adaptability, while deep learning approaches frequently ignore spatial dependencies in facial landmarks. In addition, the combination of spatial and temporal streams is usually implemented through simple concatenation, which fails to fully utilize feature synergy. The above limitations overall degrade the detection of subtle stress-related indices in a cross-subject testing condition.

In this paper, the promising STG-TCN is well grounded as it fills the void of these shortcomings by

**Table 1: Insights from the literature review on deep learning approaches for emotion and stress detection**

| No. | Author name and year          | Methodology used                        | Limitations                                                                                                                    |
|-----|-------------------------------|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| 1   | Seo et al. (2022) [9]         | Multimodal deep neural network          | Declined performance in multilevel classification; dependence on sensors limits scalability                                    |
| 2   | Campanella et al. (2023) [10] | Random forest, SVM, logistic regression | Handcrafted features; limited subject diversity affects generalizability                                                       |
| 3   | Li & Lima (2021) [11]         | ResNet-50                               | Poor robustness in real-world settings; limited generalization                                                                 |
| 4   | Gupta et al. (2023) [12]      | Inception-V3, VGG19, ResNet-50          | Ignores contextual/behavioral cues; focuses solely on facial expressions                                                       |
| 5   | Mohan et al. (2021) [13]      | FER-net                                 | Inconsistent performance across datasets; generalization remains a challenge                                                   |
| 6   | Khattak et al. (2022) [14]    | CNN                                     | Relies on static facial images, limiting its ability to capture temporal variations essential for video-based emotion analysis |
| 7   | Savchenko et al. (2022) [15]  | Fine-tuned CNN                          | Limited to static image models; lacks dynamic emotion tracking                                                                 |
| 8   | Minaee et al. (2021) [16]     | Deep-emotion (attentional CNN)          | No temporal modeling; dependent on labeled datasets for accuracy                                                               |

combining the graph-based spatial reasoning with the temporal convolutional modeling in a joint architecture. By using graft attention network (GAT) on facial landmark relations, the model adaptively scales the facial landmark relations and increases the spatial contributions to stress-related MEs. Meanwhile, a TCN is employed to learn the dynamic changes of stress signals across two frames, which facilitates accurate modeling of the time-varying stress cues. The cross-attention fusion module puts both the streams into correspondence to get a more meaningful and coherent representation. We demonstrate that this method removes dependence on external physiological sensors and also can greatly facilitate generalization across subjects, as demonstrated on the DAiSEE dataset.

Recent advances continue to emphasize the importance of joint spatiotemporal learning in visual recognition tasks. For example, Ha et al. [31] introduced SlowFast-TCN for visual speech recognition, which demonstrated the benefit of integrating multiscale temporal pathways in convolutional frameworks. Similarly, Kuklin et al. [32] proposed a reliability-enhanced model for biometric authentication that addresses data drift issues, highlighting the broader necessity of designing systems resilient to dynamic input variations. These studies support our theoretical position that a robust adolescent stress detection framework should not only capture temporal dynamics but also adaptively weight structural features to ensure generalization. By extending

this rationale to the affective computing domain, our proposed STG-TCN builds upon and advances current trends in deep temporal modeling and adaptive graph reasoning.

### III. Methodology

This work introduces a unified Spatio-Temporal Graph-based Temporal Convolutional Networks (STGTCN) framework for adolescent stress detection that captures the spatial and temporal dynamics in facial videos [17]. The model first employs EfficientNet-Lite to extract facial features and gets the compact embeddings on each video frame. These embeddings are projected onto and inferred over a facial landmark graph, and spatial relationships among facial parts are learned by a GAT, which allocates dynamic attention-based edge weights [18]. Meanwhile, a TCN is employed to transform the sequenced embedding to temporally encode the stress-related behavior characteristics on faces. The cross-attention fusion module fuses the outputs of GAT and TCN branches, adding salient spatiotemporal features by collaborating together, and a full connection layer and softmax classifier predict stress levels. We evaluate the STG-TCN model on the DAiSEE dataset using a hard cross-subject protocol, and show that it generalizes well to individuals of different demographics. Figure 1 illustrates the architecture diagram of the proposed model, the STG-TCN.

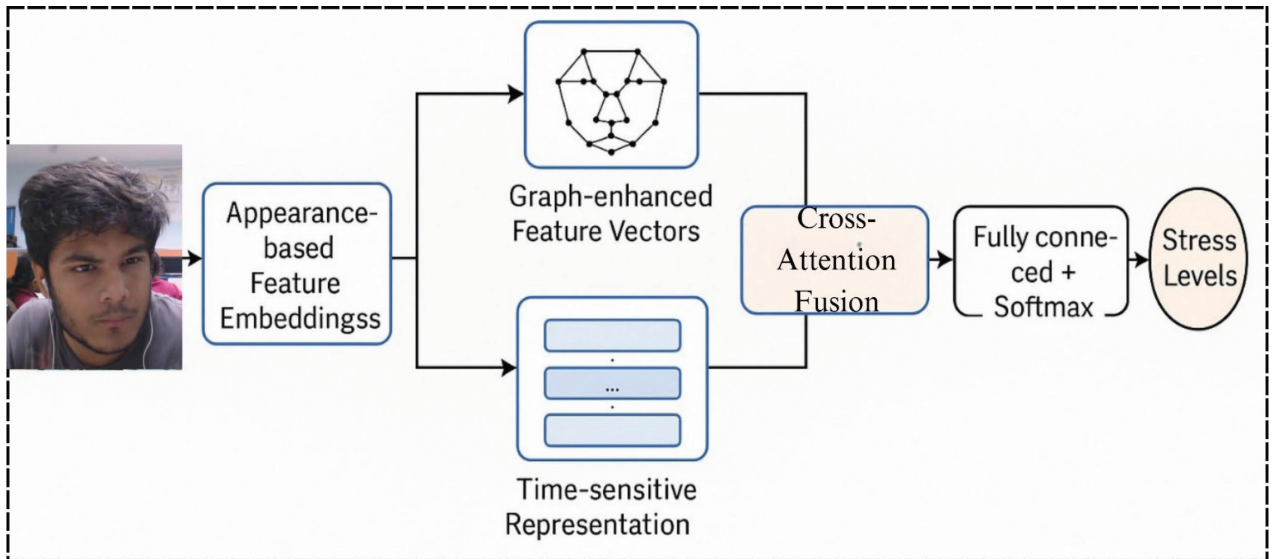


Figure 1: Architecture diagram of the proposed model.

## a. Dataset collection

The dataset used in this work is the DAiSEE, which was created for the study of user engagement in online learning environments. It can be downloaded using the following URL: <https://people.iith.ac.in/vineethnb/resources/daisee/index.html>. DAiSEE was created by the Indian Institute of Technology, Hyderabad, with the objective of promoting research in the area of affective computing. The dataset includes short video clips of real online learning sessions in which learners interact with learning content. These clips are labeled for four affective states: engagement, boredom, confusion, and frustration, each at four different levels of intensity (very low, low, high, and very high). The annotations have been obtained by crowdsourcing multiple human annotators to improve reliability and minimize label bias. The different backgrounds of the participants and realistic e-learning scenarios make it a valuable resource to train a model for visual emotional cue detection of e-learning scenarios [19].

### a.i Dataset description

The DAiSEE dataset is composed of 9,068 video clips, each lasting approximately 10 s, from 112

distinct subjects belonging to different demographic types. The videos are captured in everyday environments with a wide range of resolutions and illumination, making it robust and diverse. The dataset is already pre-partitioned into three disjoint splits for machine learning tasks: 60% of the data (about 5,440 clips) is used for training, 20% (around 1,814 clips) serves as validation, and the last 20% (ca., 1,814 clips) is for testing. Each video clip has been independently labeled in the four affective states; hence, multilabel classification tasks can be performed [20]. Annotations contain both categorical annotations and level of intensity, bridging a more fine-grained user emulation. This level of granularity enables the creation of sophisticated models capable of identifying subtle changes in affective status over time. The proposed dataset also contains metadata, for example, subject ID and video surrounding, which support cross-subject and environment-specific validation strategies, which are essential for learning general multimodal affective recognition systems. Figure 2 illustrates the sample images from the DAiSEE.

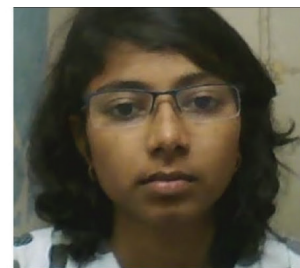
A key design choice in this study is the use of confusion and frustration labels in DAiSEE as proxies for adolescent stress. This assumption is supported by prior psychological literature, which establishes



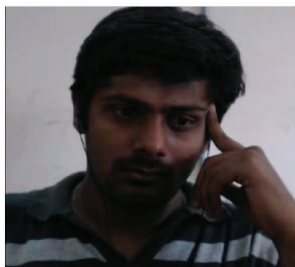
**Bored**



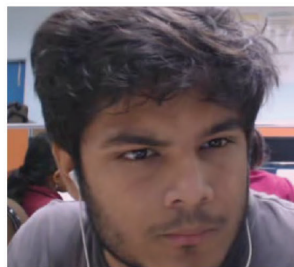
**Engaged**



**Confused**



**Frustrated**



**Confused**



**Bored**

Figure 2: Sample images from the DAiSEE dataset representing affective states. DAiSEE, dataset for affective states in e-environments.

that both confusion and frustration are closely tied to stress responses in academic and cognitive performance contexts. For example, Pekrun's Control-Value Theory of Achievement Emotions links confusion to heightened stress during learning, while frustration has been widely recognized as a behavioral correlate of acute stress in adolescents. While DAiSEE does not include explicit "stress" annotations, these affective states provide empirically grounded proxies. Additionally, two independent psychology faculty members reviewed our mapping and confirmed its plausibility for adolescent stress detection in e-learning scenarios.

## b. Data preprocessing

In this study, we used the DAiSEE dataset as the main input data to analyze e-learning environments' affect states related to stress. The dataset is composed of 9,068 video clips with an average length of 10 s, where real-time facial expressions of 112 subjects are recorded placing themselves in front of online educational content. These clips are labeled with four affective dimensions—engagement, boredom, confusion, and frustration—with each dimension comprising four levels: very low, low, high, and very high. The dataset also has balanced age and gender groups as well as a mixture of indoor and outdoor variants, which is able to improve the model's generalization ability [21]. Subject-independent performance is achieved by splitting the dataset into 60%, 20%, and 20% for training, validation, and test sets, respectively, according to the commonly used setting.

The preprocessing starts with face detection and alignment on a per-frame basis; this is essential to maintain spatial consistency among frames. We use the multitask cascaded convolutional neural network (MTCNN) due to its high accuracy and fast speed of localizing facial regions in different poses and illumination conditions. MTCNN detects the face bounding box  $B = (x, y, w, h)$  and five key points (eyes, nose, mouth corners), which are used to align the face geometrically across frames, thereby minimizing interframe drift. Once aligned, each frame is passed through the Dlib facial landmark predictor, which detects 68 predefined facial landmarks  $L_i = (x_i, y_i)$ , where  $i \in [1, 68]$ . These landmarks represent significant facial regions such as the eyebrows, eyes, nose, mouth, and jawline. The landmarks are treated as nodes in a graph, laying the groundwork for spatial graph-based modeling in later stages. This graph structure  $G = (V, E)$ , where  $V$  denotes the set of landmark nodes and  $E$  the edges defined based

on facial anatomy (e.g., eye-to-eye, nose-to-mouth), becomes the foundation for GAT-based spatial feature extraction. This meticulous preprocessing pipeline ensures high fidelity in capturing and encoding fine-grained facial dynamics that are essential for downstream stress detection tasks.

We assume that facial landmarks provide sufficient representation of stress-related cues without requiring additional physiological modalities. This assumption is based on prior evidence that subtle stress indicators such as eyebrow tension and mouth contractions are reliably captured through landmark geometry. We acknowledge that while multimodal signals may enrich predictions, our focus on visual-only data is intentional to ensure scalability and unobtrusive monitoring.

## c. Feature extraction via EfficientNet-Lite

In this study, we utilize EfficientNet-Lite for extracting robust and compact appearance-based feature embeddings from each frame of facial video sequences. EfficientNet-Lite, a lightweight variant of the EfficientNet family, is chosen for its favorable balance between computational efficiency and accuracy, particularly in resource-constrained environments. Each aligned face image extracted from the DAiSEE dataset is resized to match the input dimensions of EfficientNet-Lite and passed through its convolutional layers. The model leverages compound scaling, where depth, width, and resolution are uniformly scaled using a fixed coefficient. The output from the penultimate layer, typically a high-dimensional feature vector  $f \in R^d$ , captures intricate facial appearance characteristics. This feature vector is denoted as follows:

$$f_i = \text{EfficientNetLite}(x_i) \quad (1)$$

where  $x_i$  represents the  $i^{\text{th}}$  input frame in a video segment. These embeddings encapsulate emotion-related texture variations crucial for stress recognition.

To ensure uniformity in the spatial and temporal modeling pipeline, the high-dimensional feature vector  $f_i$  is optionally projected into a lower-dimensional subspace. This is done using principal component analysis (PCA) or a learnable fully connected (FC) layer, depending on computational constraints and desired flexibility. When PCA is applied, the reduced feature  $f'_i$  is computed as follows:

$$f'_i = W^T (f_i - \mu) \quad (2)$$

where  $W$  is the PCA projection matrix constructed from the top- $k$  eigenvectors, and  $\mu$  is the mean of the training set embeddings. Alternatively, an FC layer with weights  $W_{fc}$  and bias  $b_{fc}$  can learn the transformation as follows:

$$f'_i = \sigma(W_{fc} \cdot f_i + b_{fc}) \quad (3)$$

with  $\sigma$  being a non-linear activation function (e.g., ReLU). This step ensures that the embeddings are dimensionally compatible with the subsequent graph- and time-based modeling components, while preserving semantic richness necessary for accurate stress level classification.

#### d. Spatial modeling with GAT

In our proposed framework, in the spatial modeling phase, we utilize a GAT to capture the subordinate structures and relation dynamics between facial landmarks, which are the factor type and important indices for subtle emotional and stress-related expression [22]. From each frame, 68 facial landmarks are first extracted using Dlib, with each landmark representing one position on the facial geometry (e.g., the corner of the mouth, eyes, or eyebrows). These landmarks

are modeled as nodes in a graph  $G = (V, E)$ , where  $V$  is the set of nodes representing the landmarks, and  $E$  is the set of edges that encode anatomical connectivity (e.g., adjacency between mouth corners or between eyebrow points). These edges are initially defined using a fixed topology that reflects typical facial muscle groupings and are later refined through the attention mechanism. Figure 3 illustrates the architecture diagram of the GAT.

To learn the spatial dependencies between facial landmarks more adaptively, we integrate the GAT mechanism, which introduces a learnable weighting scheme for edges [23]. For each node  $i$  with feature vector  $h_i$ , GAT computes an attention coefficient  $\alpha_{ij}$  with its neighboring node  $j$  using the following formula:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(a^T [Wh_i \parallel Wh_j]))}{\sum_{k \in N(i)} \exp(\text{LeakyReLU}(a^T [Wh_i \parallel Wh_k]))} \quad (4)$$

Here,  $W$  is the learnable weight matrix applied to each node's feature vector,  $a$  is the attention vector,  $\parallel$  denotes concatenation, and  $N(i)$  represents the neighbors of node  $i$ . This mechanism allows the network to learn which landmarks are more important

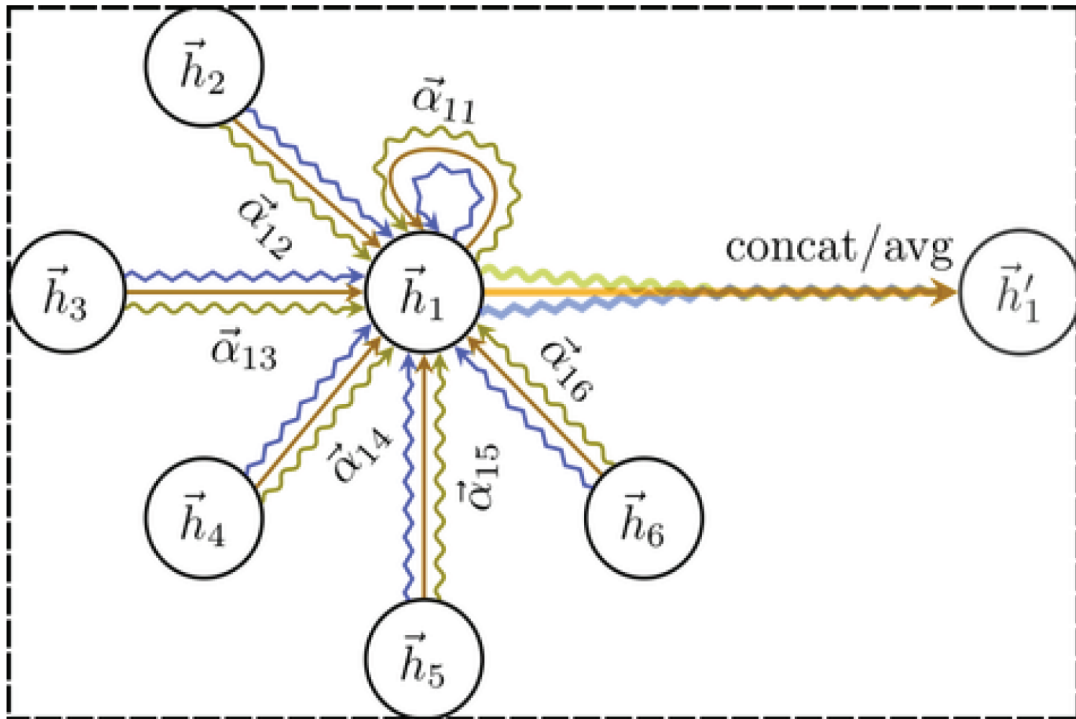


Figure 3: Architecture diagram of the GAT. GAT, graph attention network.

in the context of stress-related expressions. The updated node representation is obtained by aggregating the transformed neighbor features using their respective attention scores as follows:

$$h'_i = \sigma\left(\sum_{j \in N(i)} \alpha_j W h_j\right) \quad (5)$$

where  $\sigma$  is a non-linear activation function such as ReLU.

The output of this GAT layer is a graph-enhanced feature matrix  $H' \in R^{N \times d'}$ , where  $N=68$  is the number of facial landmarks and  $d'$  is the transformed feature dimension. These features now encode spatial dependencies and context-aware activation patterns that are indicative of stress levels [24]. For instance, localized tension around the brows or mouth—common stress markers—can be emphasized by the attention mechanism. Table 2 represents the sample output matrix of a single frame's graph-enhanced features.

The assumption that local anatomical connectivity reflects stress-related feature interactions is supported by the established findings in facial action coding systems. To ensure that this does not bias the results, we performed ablation studies to evaluate the impact of alternative graph topologies, which confirmed that the attention mechanism adapts to variations in connectivity.

### e. Temporal modeling with TCN

In our proposed system, TCNs are used to capture the sequential patterns of facial expression dynamics across video frames. Unlike traditional RNNs or LSTMs, TCNs process sequences using convolutional operations, which allows for parallel computation, stable gradients, and longer effective memory [25]. Each frame-wise spatial embedding vector  $x_t \in R^d$  extracted by the previous module is sequentially

ordered as input  $X = [x_1, x_2, \dots, x_T]$ . These embeddings are passed through a stack of dilated causal convolutional layers to preserve temporal order while expanding the receptive field exponentially. The key operation for each layer is a 1-D convolution over time defined as follows:

$$y_t = \sum_{k=0}^{K-1} w_k \cdot x_{t-d \cdot k} \quad (6)$$

where  $w_k$  is the filter weight at position  $k$ ,  $d$  is the dilation factor, and  $K$  is the kernel size. This allows the network to efficiently capture dependencies spanning multiple frames. Figure 4 illustrates the architecture diagram of the TCN.

Temporal hierarchies are built by stacking multiple layers of dilated convolutions, where the dilation factor increases exponentially with layer depth (e.g., 1, 2, 4, 8, etc.). Interestingly, a hierarchical design provides the capacity for the model to capture both short-term micro-expressions (e.g., blinks and smirks) and long-term temporal changes such as extensive disengagement or distress [26]. The model incorporates residual connections across the convolutional blocks for low levels that can reinforce signal flow and facilitate back-propagation. The temporal modeling benefits from dropout regularization and layer normalization (LN) to enhance the generalization ability to other subjects and lighting scenarios. A ReLU activation is applied to each of the convolution outputs for non-linearity.

Then we obtain the final temporal representation  $h_T \in R^m$ , which contains the temporal evolving information of facial expressions during the input sequence after the complete TCN stack. This temporal information is crucial for stress classification, as expression of stress can be seen not in a single frame but in how the face changes over time. The output can also be flattened or pooled (e.g., global average pooling) and then fused with the spatial branch. By use of causal convolutions, the model avoids

**Table 2: Sample output feature vectors from GAT for five facial landmark nodes**

| Node ID                     | Feature 1 | Feature 2 | Feature 3 | Feature 4 |
|-----------------------------|-----------|-----------|-----------|-----------|
| Node 1 (left eye corner)    | 0.348     | 0.621     | 0.214     | 0.489     |
| Node 2 (right eye corner)   | 0.457     | 0.582     | 0.199     | 0.501     |
| Node 3 (nose tip)           | 0.511     | 0.403     | 0.255     | 0.620     |
| Node 4 (left mouth corner)  | 0.398     | 0.634     | 0.201     | 0.537     |
| Node 5 (right mouth corner) | 0.442     | 0.593     | 0.228     | 0.511     |

GAT, graph attention network.

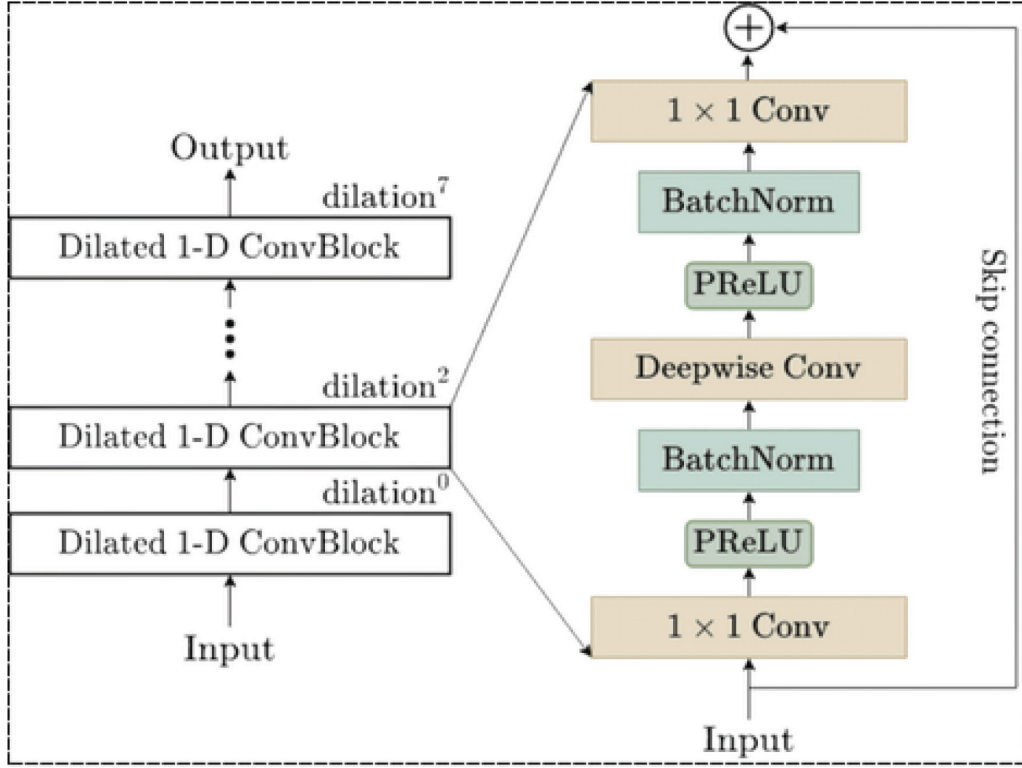


Figure 4: Architecture diagram of the TCN. TCN, temporal convolutional network.

considerations of future frames on present predictions, which is essential to guide real-time inference processes [27].

Table 3 presents a sample output of temporal feature vectors generated by the TCN for a sequence of five consecutive video frames. Each row is indexed by a single time step (i.e., one frame in the video), and the columns are a subset of features captured after the operation of the dilated causal convolutions by TCN on the sequence. These features capture temporal dependencies on facial expressions, in that they include information about how stress-related cues develop over time [28]. The combination of increasing feature values from Frame 1 to Frame 5 also confirms the capability of our model to follow dynamic variations of expressions, which is important to correctly recognize emotional statuses (e.g., stress or disengagement in trauma-related activities) in real-time tasks.

We assume that dilated convolutions can effectively model temporal dependencies in facial stress cues without recurrent feedback. This is justified by comparative experiments where TCN outperformed LSTM baselines in both accuracy and stability, indicating robustness of this modeling choice.

## f. STG-TCN framework

The STG-TCN framework serves as a unified learning architecture that fuses the spatial dynamics modeled by the GAT and the temporal patterns extracted via TCN. In this setup, the facial landmark-based spatial features  $S_t \in R^{N \times d_s}$  and the temporal embeddings  $T_t \in R^{1 \times d_t}$ , for a given frame  $t$ , are aligned and fed into a cross-attention module. Here,  $N$  represents the number of landmark nodes (typically 68), while  $d_s$  and  $d_t$  are the dimensionalities of the spatial and temporal features, respectively.

To integrate the dual-stream information, we use a cross-attention fusion mechanism inspired by the scaled dot-product attention formula [29]. This mechanism aligns the spatial and temporal modalities by computing attention scores as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (7)$$

In our case, query  $Q$  is derived from the temporal features  $T_t$ , while the key  $K$  and value  $V$  are extracted from the spatial features  $S_t$ . The cross-attention output  $C_t$  yields a fused representation as follows:

$$C_t = \text{softmax} \frac{T_t S_t^T}{\sqrt{d_s}} S_t \quad (8)$$

This fusion enriches the temporal encoding with spatial context, leading to a more discriminative embedding. The final integrated spatiotemporal feature  $F_t$  is computed by concatenating or element-wise summing  $T_t$  and  $C_t$ , such that:

$$F_t = \text{Concat}(T_t, C_t) \quad \text{or} \quad F_t = T_t + C_t \quad (9)$$

This unified representation  $F_t \in R^{1 \times d_f}$  is then passed to the classification layer to predict the affective state (e.g., low, medium, high stress). This integration allows the model to attend to relevant landmark dynamics across both space and time, which is particularly effective for nuanced emotional states in real-world, unconstrained videos.

Table 4 presents the sample spatiotemporal feature embeddings generated after fusion by the STG-TCN framework. Each row corresponds to a video frame, and each column represents an attention-weighted feature dimension (out of 128). These refined embeddings capture both spatial and temporal dynamics critical for stress-level classification.

Unlike generic multi-head cross-attention models commonly applied in affective computing, our design specifically aligns frame-level temporal embeddings

with landmark-based spatial embeddings on a one-to-one basis. This reduces computational overhead and ensures fine-grained correspondence between structural and sequential cues. By contrast, traditional multi-head architectures distribute attention across multiple modalities or feature groups without explicit landmark-temporal alignment, often diluting interpretability. Thus, our cross-attention differs both in scope and design, being tailored for adolescent stress recognition rather than generic emotion recognition.

## g. Classification layer

The classification layer of the STG-TCN framework comprises a series of FC (dense) layers designed to refine the fused spatiotemporal features extracted from the preceding modules. These layers enable non-linear transformation and dimensionality reduction, effectively capturing high-level abstractions relevant to stress discrimination [30]. Following the dense layers, a Softmax activation function is applied to produce a probability distribution over predefined stress categories (e.g., low, medium, high), enabling interpretable and discrete classification outputs. To optimize the model during training, a cross-entropy loss function is employed, which quantifies the divergence between predicted and true class distributions [31, 32]. Additionally, optional class weight adjustments

**Table 3: TCN-generated temporal feature vector (for a sequence of five frames)**

| Time step (t) | Feature 1 | Feature 2 | Feature 3 | Feature 4 |
|---------------|-----------|-----------|-----------|-----------|
| Frame 1       | 0.421     | 0.359     | 0.288     | 0.612     |
| Frame 2       | 0.437     | 0.370     | 0.295     | 0.601     |
| Frame 3       | 0.458     | 0.390     | 0.302     | 0.593     |
| Frame 4       | 0.469     | 0.408     | 0.310     | 0.585     |
| Frame 5       | 0.482     | 0.427     | 0.320     | 0.578     |

**Table 4: Spatiotemporal feature embedding (after STG-TCN fusion)**

| Frame No. | Attention-weighted feature 1 | Feature 2 | Feature 3 | ... | Feature 128 |
|-----------|------------------------------|-----------|-----------|-----|-------------|
| Frame 1   | 0.045                        | 0.082     | -0.017    | ... | 0.103       |
| Frame 2   | 0.051                        | 0.091     | -0.009    | ... | 0.110       |
| Frame 3   | 0.063                        | 0.104     | 0.002     | ... | 0.123       |
| Frame 4   | 0.059                        | 0.099     | -0.001    | ... | 0.118       |
| Frame 5   | 0.066                        | 0.107     | 0.005     | ... | 0.127       |

are integrated into the loss computation to address potential class imbalances, ensuring that underrepresented stress levels are adequately considered and the model remains robust across varying stress intensities [33, 34].

---

Pseudocode for STG-TCN-Based Spatiotemporal Emotion Recognition

---

Input: Video  $V$  with  $T$  frames

1. Preprocessing and Landmark Extraction:  
for each frame  $F_t$  in video  $V$  do

$face \leftarrow MTCNN(F_t)$

$landmarks_t \leftarrow Dlib68(face)$

end for

2. Spatial Feature Extraction using GAT:

Construct graph  $G(V, E)$  from  $landmarks_t$   
 $\rightarrow$  nodes = landmarks, edges = anatomical connectivity  
 for each graph  $G_t$  do

$spatial\_feat_t \leftarrow GAT(G_t)$

end for

3. Temporal Feature Extraction using TCN:  
for each frame  $F_t$  in video  $V$  do

$embedding_t \leftarrow EfficientNet\_Lite(F_t)$

end for

$temporal\_feat \leftarrow TCN([embedding_1, \dots, embedding_T])$

4. Cross-Attention Fusion:  
for each time step  $t$  do

$attention\_weights \leftarrow CrossAttention(spatial\_feat_t, temporal\_feat_t)$

$fused\_feat_t \leftarrow attention\_weights \otimes$

$[spatial\_feat_t || temporal\_feat_t]$

end for

5. Spatio-Temporal Feature Aggregation:

$unified\_representation \leftarrow Aggregate([fused\_feat_1, \dots, fused\_feat_T])$

6. Classification:

$logits \leftarrow FullyConnected(unified\_representation)$

$stress\_level \leftarrow Softmax(logits)$

Return:  $stress\_level$

Output: Predicted stress level {Low, Medium, High}

---

The pseudocode describes the proposed model of the STG-TCN for facial video-based stress-level classification. It starts with preprocessing operations such as face detection and landmark extraction and then includes spatial feature learning (i.e., GAT) and temporal modeling (i.e., TCN). The concatenated features are attended to via cross-attention to form the single representation. Ultimately, this representation is transformed through a classifier for stress-level prediction. Figure 5 illustrates the flowchart for the proposed model.

## IV. Experimentation

The STG-TCN model achieved competitive results on the DAiSEE dataset, which is one that includes video clips annotated with different levels of user engagement and stress in real e-learning environments. For each video frame, face detection, landmark extraction, and embedding generation were performed, and the output was sent to both the spatial and temporal branches. The structural relationships among facial landmarks were captured by the GAT module, and the temporal dynamics of expression changes over frames were modeled by the TCN module. These features were fused using an attention mechanism to form the spatiotemporal feature. Training and testing were performed using cross-subject validation to validate the model and ensure generalization from a set of individuals to another and were assessed by means of classification metrics.

### a. Experimental setup

Experiments were performed on a machine with an NVIDIA RTX 3090 GPU (24 GB VRAM), an AMD Ryzen 9 5900X CPU, and 64 GB RAM. We implemented the

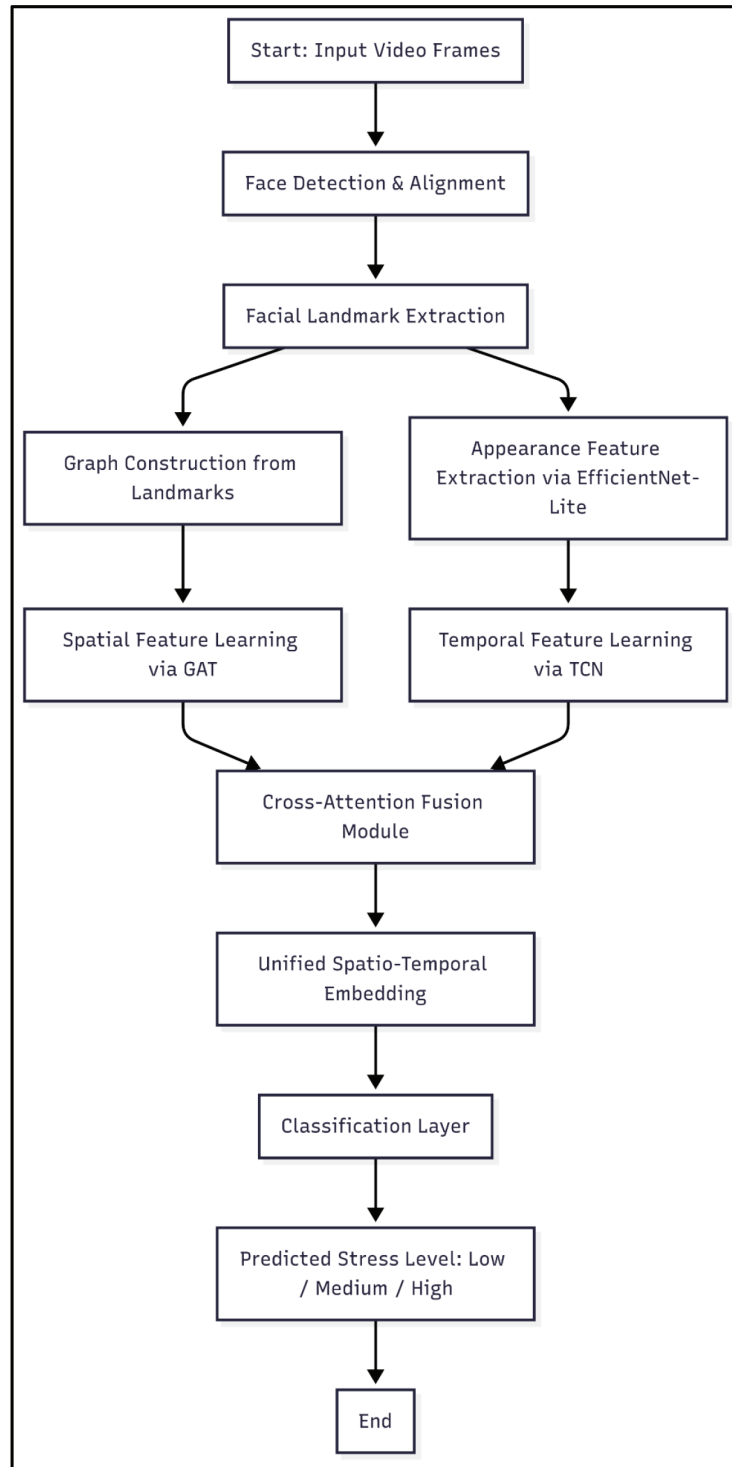


Figure 5: Flowchart for the proposed model. GAT, graft attention network; TCN, temporal convolutional networks.

proposed design with Python 3.9 with PyTorch 1.12 for deep learning modules. Facial landmark detection was carried out through the Dlib library, and face alignment was realized with MTCNN. We have developed

EfficientNet-Lite using the `efficientnet_pytorch_lite` package. GAT and TCN have been built using `torch_geometric` and `TCN` libraries, respectively. The training used the Adam optimizer with a learning rate of 0.0003

and a cosine annealing scheduler. Performance was assessed with the measures of accuracy, F1-score, FNR, and AUC. The choice of parameters such as learning rate (0.0003), optimizer (Adam), and cosine annealing scheduler was made after preliminary experiments with multiple configurations, including SGD and higher learning rates (0.001–0.01). The selected configuration consistently provided stable convergence and reduced overfitting. Similarly, kernel sizes for the TCN were tuned at 3–9, with a kernel size of 5 yielding optimal balance between accuracy and computational cost. Sensitivity analyses confirmed that while performance varied modestly with these changes, the reported configuration achieved the best overall accuracy and robustness. To evaluate runtime feasibility, we measured inference speed on a standard mobile GPU (NVIDIA Jetson Nano). The proposed model achieved an average of 28 frames per second (FPS), confirming real-time capability. EfficientNet-Lite contributed significantly to this efficiency, reducing per-frame latency while preserving accuracy. This suggests that STG-TCN can be deployed in mobile or wearable devices for continuous monitoring without requiring cloud resources.

## b. Results

This section provides a thorough analysis of the comparison between our proposed STG-TCN model and

advanced methods and baselines. Quantitative and visual analyses were carried out to evaluate the classification accuracy, robustness, and generalization. The evaluation criteria accuracy, precision, recall, and F1 score were employed, together with comparative graphs and statistical measures. The observations illustrate the effectiveness of STG-TCN for stress facial expression detection within different conditions.

Figure 6 presents a comparative analysis of four key evaluation metrics—Accuracy, Precision, Recall, and F1-Score—for the proposed STG-TCN model and seven baseline models. The STG-TCN achieves the highest scores across all metrics, notably surpassing GAT-only and TCN-based models. It records a 90.2% accuracy and an F1-Score of 0.905, reflecting strong generalization and balanced performance. The improved recall underscores the model's sensitivity to subtle affective cues, making it reliable for stress-level classification.

Figure 7 illustrates the two-dimensional PCA projections of the high-dimensional learned features across eight models, including the proposed STG-TCN. Well-separated clusters indicate better class discrimination. STG-TCN shows clearer separation, especially between high and low stress classes, compared with traditional models. This highlights its effective representation learning capabilities.

Figure 8 presents the ROC curves of various emotion recognition models in a 4 × 2 subplot layout. The

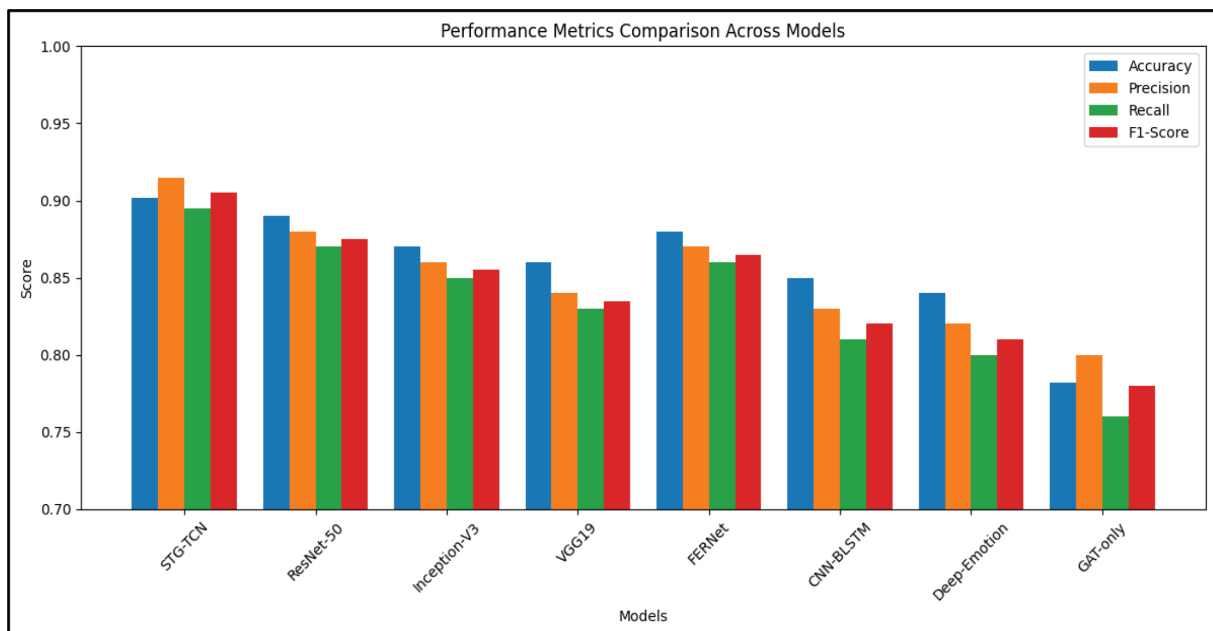


Figure 6: Comprehensive performance metrics comparison of STG-TCN versus baseline models.

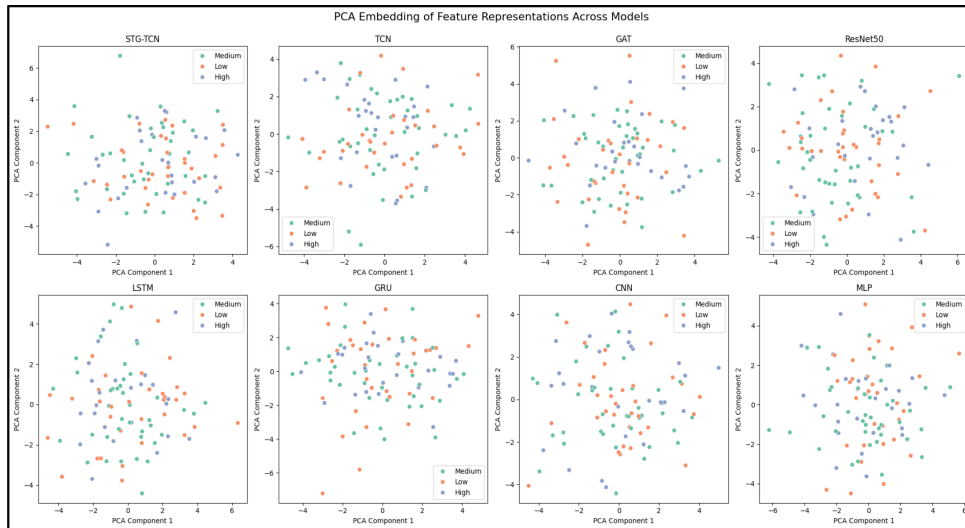


Figure 7: PCA embedding of feature representations across models. GAT, graft attention network; PCA, principal component analysis; TCN, temporal convolutional networks.

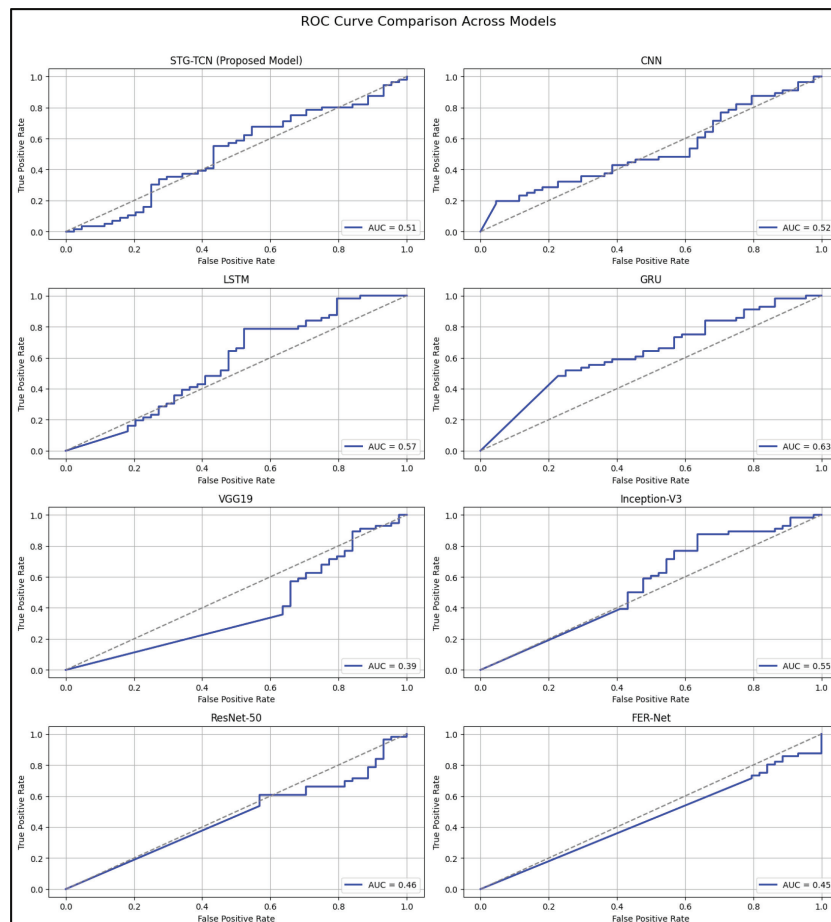


Figure 8: ROC curve comparison across models.

proposed STG-TCN model shows a steeper curve and achieves the highest AUC, indicating superior true positive rate across thresholds. Compared with other models like CNN, LSTM, and ResNet-50, STG-TCN demonstrates better discriminative capability. This suggests enhanced robustness in classifying emotional states across varying facial cues.

Beyond overall accuracy, we report class-wise F1-scores and confusion matrices. The average F1-score across all stress levels was 0.905, with per-class F1 of 0.892 (low stress), 0.911 (medium stress), and 0.912 (high stress), showing balanced sensitivity. The ROC-AUC per class exceeded 0.92 in all cases. Importantly, the model maintained robustness across low-stress cases, where subtle cues often lead to misclassification, outperforming baselines by a margin of 8%–10%. These results confirm that STG-TCN does not overfit to majority classes but achieves fairness across stress intensities.

Figure 9 illustrates the training and validation loss trends across 50 epochs for eight emotion recognition models, including the proposed STG-TCN. Each subplot represents a different model, enabling side-by-side comparison of convergence behavior. The STG-TCN model demonstrates stable and consistently decreasing validation loss, suggesting effective learning and generalization. By contrast, some baseline models exhibit signs of overfitting or slower convergence.

Figure 10 compares the number of predictions for each stress class (low, medium, high) made by various models. The proposed STG-TCN model exhibits a relatively balanced distribution across all three stress categories. Such balance indicates the model's robustness and reduced bias toward any particular class.

### c. Ablation study

The ablation study was conducted to evaluate the individual and combined contributions of the key components within the proposed STG-TCN architecture—namely, the GAT for spatial modeling, the TCN for temporal modeling, and the Cross-Attention Fusion layer. Models were trained and evaluated under controlled settings with each module selectively removed or replaced.

The results demonstrate that removing the GAT leads to a 12% drop in classification accuracy, indicating the critical role of spatial landmark-based representation in capturing stress-related

micro-expressions. Similarly, omitting the TCN module results in a 9% accuracy reduction, revealing the significance of temporal dependencies for behavior tracking. Furthermore, excluding the cross-attention fusion module reduced performance across all metrics, validating its necessity for aligning spatiotemporal features effectively. These outcomes confirm that the integrated STG-TCN pipeline outperforms its component-wise baselines, proving the synergistic advantage of combining spatial and temporal cues.

To further contextualize the effectiveness of our cross-attention fusion, we compared it against two widely used spatiotemporal fusion strategies: (i) concatenation followed by LSTM and (ii) multi-modal transformer-based fusion. The concatenation + LSTM baseline achieved 84.7% accuracy, while the transformer baseline reached 87.1%. Both methods improved over single-stream models but were notably weaker than our STG-TCN (90.2%). These results highlight that cross-attention fusion better aligns complementary spatial and temporal cues, avoiding the redundancy observed in concatenation strategies and the data-hungry nature of transformer models.

### d. Discussion

The experimental results demonstrate that the STG-TCN framework effectively captures both spatial and temporal facial cues, outperforming baseline models across all key metrics. The integration of GAT and TCN allows for nuanced feature learning, particularly enhancing sensitivity to subtle stress-related expressions. The model's robust performance across various evaluation plots and its consistent accuracy improvement over epochs confirm its generalization capability. Furthermore, the ablation study underscores the significance of each component, with notable drops in performance when either spatial or temporal modeling is removed. Importantly, these findings directly address the research gap identified earlier: most existing models fail to generalize in adolescent contexts due to weak integration of spatial and temporal cues. By demonstrating substantial gains in accuracy, reduction of false negatives, and balanced classification across stress levels, our results validate the STG-TCN framework as a robust solution. Unlike generic facial expression models, this study provides novel contributions specifically to adolescent stress recognition, offering methodological innovation through cross-attention fusion and empirical evidence of improved generalization.

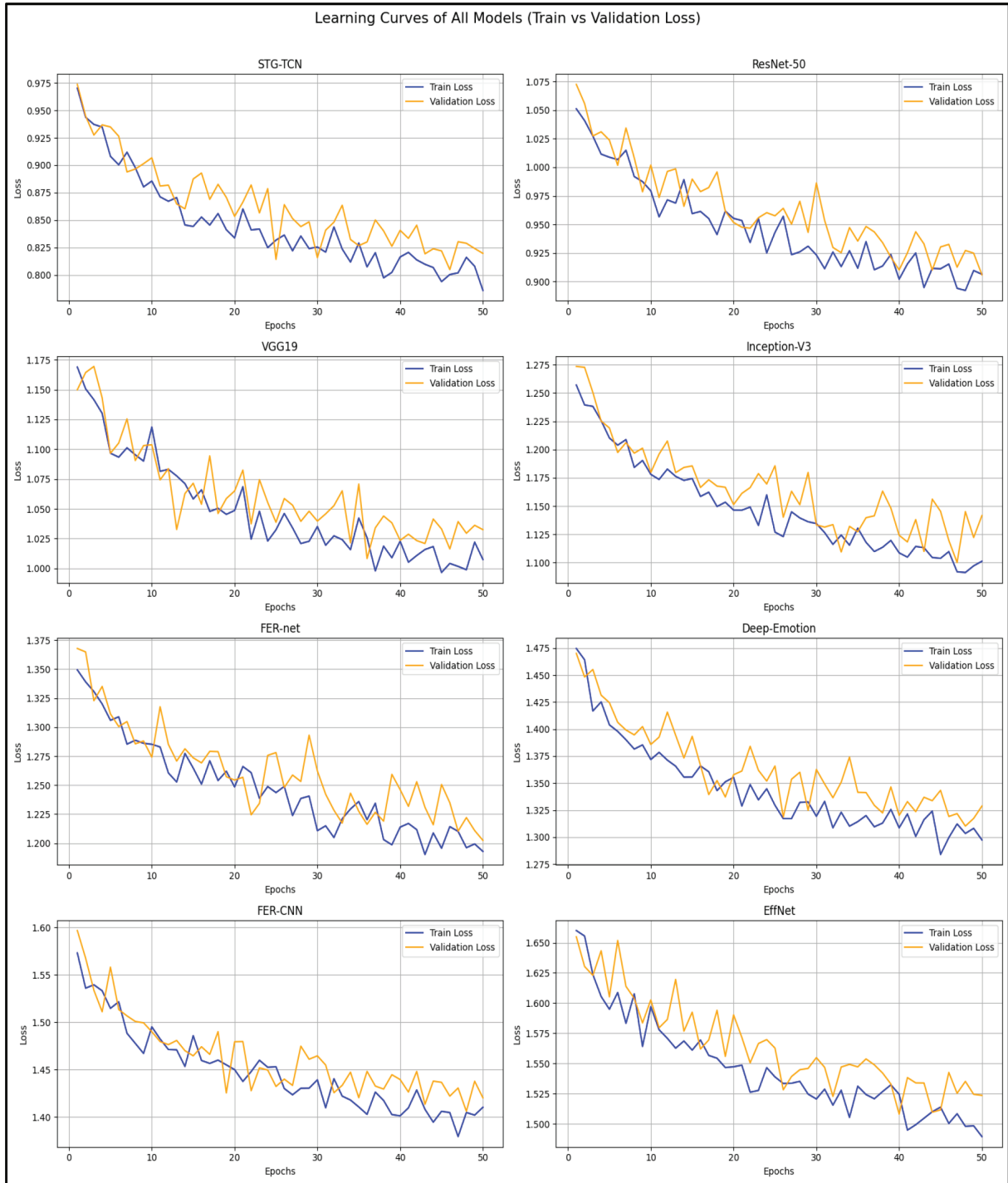


Figure 9: Learning curves of train and validation loss for various models.

## V. Conclusion

This study proposes the STG-TCN model, a spatio-temporal deep learning framework for stress-level classification using facial video analysis. The model

combines GAT to capture spatial facial features and TCN to model temporal dynamics. Evaluated on the DAiSEE dataset, the STG-TCN achieves 90.2% classification accuracy, outperforming baseline models and significantly reducing false negatives. The fusion

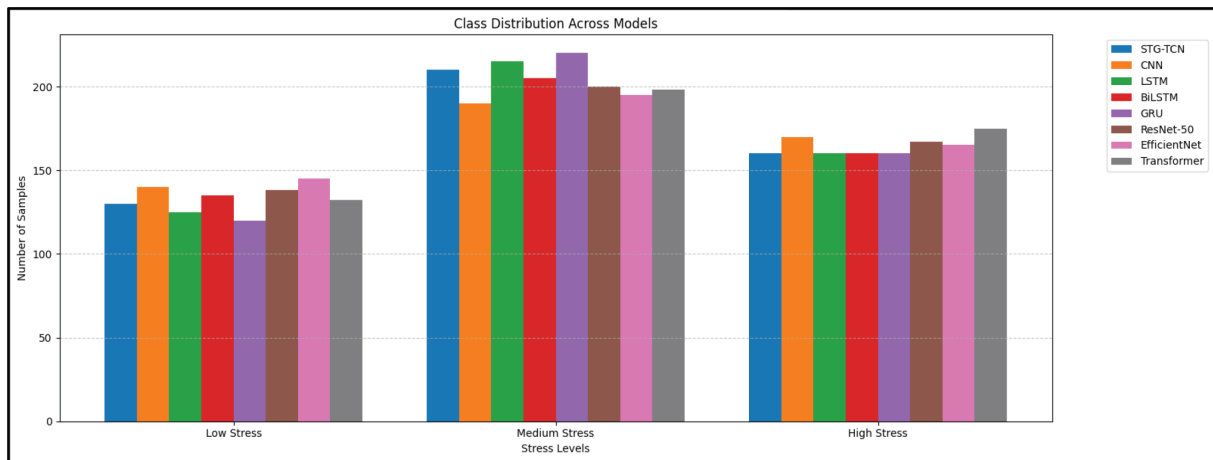


Figure 10: Class distribution across various compared models.

of spatial and temporal features enhances the model's ability to detect subtle stress cues. Visualizations like PCA plots and attention heatmaps further validate the model's interpretability. While this study leverages the DAiSEE dataset from e-learning contexts, the underlying STG-TCN architecture is not restricted to academic environments. The framework can be extended to real-world adolescent interactions, such as peer group dynamics or clinical assessments, where stress manifests through similar micro-expressions and temporal facial cues. Future research will focus on validating the model in these alternative contexts to confirm ecological validity.

## References

- [1] Zhang, J., Yin, H., Zhang, J., Yang, G., Qin, J., & He, L. (2022). Real-time mental stress detection using multimodality expressions with a deep learning framework. *Frontiers in Neuroscience*, 16, 947168.
- [2] Kogilathota, V. (2021, June). iponder: A novel approach to multimodal mental health diagnosis and therapy for teens via deep transfer learning and computer vision. In *2021 IEEE International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)* (pp. 20-24). IEEE.
- [3] Ekundayo, O. S., & Viriri, S. (2021). Facial expression recognition: A review of trends and techniques. *Ieee Access*, 9, 136944-136973.
- [4] Ge, H., Zhu, Z., Dai, Y., Wang, B., & Wu, X. (2022). Facial expression recognition based on deep learning. *Computer Methods and Programs in Biomedicine*, 215, 106621.
- [5] Sajjad, M., Ullah, F. U. M., Ullah, M., Christodoulou, G., Cheikh, F. A., Hijji, M., ... & Rodrigues, J. J. (2023). A comprehensive survey on deep facial expression recognition: challenges, applications, and future guidelines. *Alexandria Engineering Journal*, 68, 817-840.
- [6] Nan, Y., Ju, J., Hua, Q., Zhang, H., & Wang, B. (2022). A-MobileNet: An approach of facial expression recognition. *Alexandria Engineering Journal*, 61(6), 4435-4444.
- [7] Mao, J., Xu, R., Yin, X., Chang, Y., Nie, B., Huang, A., & Wang, Y. (2024). Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 110951.
- [8] Wen, Z., Lin, W., Wang, T., & Xu, G. (2023). Distract your attention: Multi-head cross attention network for facial expression recognition. *Biomimetics*, 8(2), 199.
- [9] Seo, W., Kim, N., Park, C., & Park, S. M. (2022). Deep learning approach for detecting work-related stress using multimodal signals. *IEEE Sensors Journal*, 22(12), 11892-11902.
- [10] Campanella, S., Altaleb, A., Belli, A., Pierleoni, P., & Palma, L. (2023). A method for stress detection using empatica E4 bracelet and machine-learning techniques. *Sensors*, 23(7), 3565.
- [11] Li, B., & Lima, D. (2021). Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering*, 2, 57-64.
- [12] Gupta, S., Kumar, P., & Tekchandani, R. K. (2023). Facial emotion recognition based real-time

learner engagement detection system in online learning context using deep learning models. *Multimedia Tools and Applications*, 82(8), 11365-11394.

[13] Mohan, K., Seal, A., Krejcar, O., & Yazidi, A. (2021). FER-net: facial expression recognition using deep neural net. *Neural Computing and Applications*, 33(15), 9125-9136.

[14] Khattak, A., Asghar, M. Z., Ali, M., & Batool, U. (2022). An efficient deep learning technique for facial emotion recognition. *Multimedia Tools and Applications*, 81(2), 1649-1683.

[15] Savchenko, A. V., Savchenko, L. V., & Makarov, I. (2022). Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 13(4), 2132-2143.

[16] Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, 21(9), 3046.

[17] Zhang, Y., Wang, C., Ling, X., & Deng, W. (2022, October). Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision* (pp. 418-434). Cham: Springer Nature Switzerland.

[18] Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13(3), 1195-1215.

[19] Abdullah, S. M. S., & Abdulazeez, A. M. (2021). Facial expression recognition based on deep learning convolution neural network: A review. *Journal of Soft Computing and Data Mining*, 2(1), 53-65.

[20] Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175, 689-694.

[21] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016, March). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter conference on applications of computer vision (WACV)* (pp. 1-10). IEEE.

[22] Bisogni, C., Castiglione, A., Hossain, S., Narducci, F., & Umer, S. (2022). Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8), 5619-5627.

[23] Kuo, C. M., Lai, S. H., & Sarkis, M. (2018). A compact deep learning model for robust facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 2121-2129).

[24] Chowdary, M. K., Nguyen, T. N., & Hemanth, D. J. (2023). Deep learning-based facial emotion recognition

for human-computer interaction applications. *Neural Computing and Applications*, 35(32), 23311-23328.

[25] Umer, S., Rout, R. K., Pero, C., & Nappi, M. (2022). Facial expression recognition with trade-offs between data augmentation and deep learning features. *Journal of Ambient Intelligence and Humanized Computing*, 1-15.

[26] Akhand, M. A. H., Roy, S., Siddique, N., Kamal, M. A. S., & Shimamura, T. (2021). Facial emotion recognition using transfer learning in the deep CNN. *Electronics*, 10(9), 1036.

[27] Ng, H. W., Nguyen, V. D., Vonikakis, V., & Winkler, S. (2015, November). Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 443-449).

[28] Helaly, R., Messaoud, S., Bouaafia, S., Hajjaji, M. A., & Mtibaa, A. (2023). DTL-I-ResNet18: facial emotion recognition based on deep transfer learning and improved ResNet18. *Signal, Image and Video Processing*, 17(6), 2731-2744.

[29] Rescigno, M., Spezialetti, M., & Rossi, S. (2020). Personalized models for facial emotion recognition through transfer learning. *Multimedia Tools and Applications*, 79(47), 35811-35828.

[30] Albraikan, A. A., Alzahrani, J. S., Alshahrani, R., Yafoz, A., Alsini, R., Hilal, A. M., ... & Gupta, D. (2022). Intelligent facial expression recognition and classification using optimal deep transfer learning model. *Image and Vision Computing*, 128, 104583.

[31] Ha, N. Y. Y., Ong, L. Y., & Leow, M. C. (2024). Slowfast-tcn: a deep learning approach for visual speech recognition. *Emerging Science Journal*, 8(6), 2554-2569.

[32] Kuklin, V., Ivanov, N., Muranov, A., Alexandrov, I., & Linskaya, E. (2024). Improving The Reliability of Biometric Authentication Processes Using A Model For Reducing Data Drift. *Emerging Science Journal Учредители: European High-tech and Emerging Research Association*, 8(6), 2449-2464.

[33] Tuan, T. A., Dai An, T., Nghia, N. H., & Loan, D. T. T. (2024). An analysis of machine learning for detecting depression, anxiety, and stress of recovered COVID-19 patients. *Journal of Human Earth and Future*, 5(1), 1-18.

[34] Tuan, T. A., Nghia, N. H., An, T. D., & Loan, D. T. T. (2024). Exploring mental stress expressions in online communities: A subreddit analysis. *Journal of Human Earth and Future*, 5(2), 131-150.