Lightweight Spatio-Temporal Convolutional Neural Network for Audio-Visual Emotion Recognition

Su Yen Ding, Tong Boon Tang, *Senior Member, IEEE*, and Cheng-Kai Lu, *Senior Member, IEEE*

***Abstract*—Emotion recognition (ER) enhances human-computer interaction in customer service and healthcare. However, the high computational complexity of existing convolutional neural network (CNN)-based approaches limits their real-time applicability in compact audio-visual emotion recognition (AVER) systems. This study introduces a lightweight deep learning-based AVER framework, employing a 2-layer, 1D CNN for audio analysis and a 3-layer, 2D CNN for facial image processing. The spatial model (2D CNN) reduces input complexity by using grayscale images, downsizing to 64 × 64 pixels, and performing three convolutions to extract facial patches. For audio, 1D convolution on Mel-Frequency Cepstral Coefficients (MFCCs) helps preserve essential features while lowering computational demand. With small kernel size and set of optimized parameters, the proposed framework balances performance and complexity. Benchmarking on SAVEE, RAVDESS, and MEAD datasets show accuracy of 97.57%, 95.89%, and 98.57%, respectively, demonstrating its potential for resource-constrained devices. Integrated gradient analysis further reveals key dependencies on eyebrows, eyes, and mouth for facial ER, and 40 MFCCs for audio ER.**

***Index Terms*—Audio-visual emotion recognition (AVER), convolutional neural network (CNN), hardware amenable AI, mental health.**

I. Introduction

EMOTION recognition (ER) is a cutting-edge frontier in artificial intelligence (AI), aiming to enable machines to identify and interpret human emotions, thereby enhancing human-computer interaction to meet human demands and expectations [1]. This technology is applied in areas such as customer service and healthcare systems, with recent implementation in education environments [2]. Recent advances in ER technology aim to bridge the gap between human and AI interactions.

ER leverages on advanced algorithms and machine learning techniques, particularly deep learning, to analyze and interpret emotional cues. Initial ER efforts focused on unimodal inputs such as facial expressions, speech, and textual cues [3]–[5]. While these approaches showed promise, they struggled to comprehensively capture the nuances of human emotions, which are complex and correlated with multiple cues. Abdullah et al. [6] demonstrated the inefficiency of predicting emotions with single modalities.

With the increasing demand for ER technology, research now focuses on new techniques, algorithms, and model architectures to enhance accuracy and efficiency. Conventional machine learning techniques require extensive custom pre-processing and cannot automatically extract high-level features from raw data, leading to increased manual processing time. Among deep learning algorithms, convolutional neural networks (CNNs) have garnered significant attention for their ability to predict emotions [7]. However, as CNNs become more complex with deeper convolutions, the risk of overfitting increases, reducing generalization capability [8], [9].

Although transfer learning has accelerated development by leveraging pre-trained models, it is computationally expensive and difficult to deploy on devices with constrained processing power [10]. Fine-tuning these models typically requires high-performance GPUs with specialized computing structure, which, despite their training efficiency, face constraints like limited memory, high power consumption, and latency [11]. These drawbacks make such hardware impractical and costly for real-time deployment. While cost-effectiveness is less critical during development phase, it becomes pivotal when scaling for real-time applications. This work develops an optimized model that delivers state-of-the-art performance while remaining viable for resource-constrained, cost-efficient devices, ensuring a practical balance between computational efficiency and affordability for real-time applications.

To date, the trade-off between model complexity and computational requirements remains challenging as deep learning models continue to evolve. CNN models can be categorized by input dimensions. One-dimensional (1D) CNNs are effective for processing speech audio signals with lower computational complexity, making them suitable for real-time applications [12]; two-dimensional (2D) CNNs require larger datasets to avoid overfitting and are designed for processing image data. Three-dimensional (3D) CNNs excel in video recognition but demand longer training times and more computational power [13]. To address ER challenges, there is a growing focus on lightweight multimodal systems that can achieve a high accuracy while ensuring computational efficiency to make it

Received 25 May 2024; revised 9 February 2025; accepted 30 April 2025. Date of publication 2 May 2025; date of current version 3 December 2025. This work was supported by the YUTP-FRG under Grant 015LC0-395, in part by the the Ministry of Higher Education Malaysia, in part by the Higher Institutional Centre of Excellence (HICoE) Scheme, and in part by the Centre for Intelligent Signal and Imaging Research (CISIR). Recommended for acceptance by Nagarajan Ganapathy. *(Corresponding authors: Tong Boon Tang; Cheng-Kai Lu.)*

Su Yen Ding and Tong Boon Tang are with the Centre of Intelligent Signal and Imaging Research (CISIR), Universiti Teknologi PETRONAS, Bandar Seri Iskandar 32610, Malaysia (e-mail: su_22000257@utp.edu.my; tongboon.tang@utp.edu.my).

Cheng-Kai Lu is with the Department of Electrical Engineering, National Taiwan Normal University, Taipei 106308, Taiwan (e-mail: cklu@ntnu.edu.tw).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAFFC.2025.3566773, provided by the authors.

Digital Object Identifier 10.1109/TAFFC.2025.3566773

possible deployment in resource-constrained real-world scenarios. Combining 1D and 2D CNNs, known for their effectiveness in analyzing speech audio and facial cues respectively, promises a simpler architecture than a full 3D CNNs [14], making it practical for real-time audio-visual emotion recognition (AVER) systems. Emotions are inherently complex and expressed in diverse ways, necessitating multimodal fusion approaches [15]. The superiority of multimodal approaches for AVER has been demonstrated in previous studies [16].

This paper proposes an AVER model that incorporates Mel-Frequency Cepstral Coefficients (MFCCs) from speech audio and facial cues from video frames. Using a model-level fusion strategy, embeddings from each modality are concatenated before training a fusion layer to describe emotions. The proposed model is validated with three publicly available benchmark datasets: SAVEE [17], RAVDESS [18], and MEAD [19]. We hypothesize that our CNN model can achieve competitive recognition accuracy for AVER while reducing computational demands by strategically lowering the parameter count. The main contributions of our study are: -

1. Lightweight multimodal CNN architecture for AVER: The paper introduces a novel multimodal CNN architecture for AVER, incorporating MFCCs for speech audio and facial images, and offering a 99.94% reduction in parameters compared to state-of-the-arts.
2. High accuracy with reduced complexity: The model achieves accuracies of 97.57% on SAVEE, 95.89% on RAVDESS, and 98.57% on MEAD datasets while significantly reducing parameters, enabling real-time applications on resource-constrained devices.
3. Hardware-amenable spatio-temporal model: Combining a 2-layer 1D CNN for audio analysis and a 3-layer 2D CNN for visual analysis, the model with significantly reduced parameters is suitable for deployment on cost-effective platforms like NVIDIA Jetson Orin Nano without performance degradation, hence broadening AVER's real-world applications.

II. Related Work

In this section, we introduce three types of fusion available for AVER and share the pros and cons of each approach. We further review the state-of-the-arts CNN deployed for AVER approach and report the pain-points of current solutions. We also extend our literature review on types of input feature for AVER model.

A. Priori in Audio-Visual Emotion Recognition

Common fusion techniques for AVER are feature-level (early) fusion, decision-level (late) fusion, and model-level fusion [20]. Early fusion concatenates feature vectors of each modality, while late fusion cascades predictions from each modality. Model-level fusion learns a shared representation from all modalities for prediction. Early fusion is widely used [21], but suffers from time-synchrony issues like late fusion [22]. Both early and late fusion achieve accuracies of 60–80% [23], [24]. Accuracy can be boosted up to 92% with RNN and LSTM incorporation, albeit at the expense of high computational complexity [25]. Model-level fusion can further address the limitation of video sequence length by LSTM via learning representations from each modality within a shared space before prediction and attained similar level of accuracy [26]. The next question is how to transform such model-level fusion approach to be computationally effective.

B. Audio-Visual Emotion Recognition CNN Model-Level Fusion

In [27], one 4-layer 1D CNN and one bi-directional LSTM (bi-LSTM) were proposed to process speech audio with handcrafted low-level features as inputs, while a 10-layer 2D CNN inspired by the VGG model was deployed to analyze facial cues. Features from three models were concatenated and passed through three dense layers for classification, achieving a 70.24% accuracy on IEMOCAP dataset. Similarly, Aghajani used a bi-LSTM for speech audio feature extraction, then emotion recognition with a 2-layer 2D CNN and bi-LSTM for the speech audio features, and a 6-layer 2D CNN for the facial images [28]. Pre-trained YAMNet embeddings from Mel-spectrograms replaced handcrafted features, yielding an overall accuracy of 81.04% on the RAVDESS.

Buoali et al. implemented a 1-layer 2D CNN-LSTM for 80-sequence facial images and a 1-layer 1D CNN for speech audio using Mel-spectrogram, MFCC, chromagram, spectral contrast, and tonnetz features [29]. Flattened modality features were connected to a 128-neuron dense layer, with SoftMax predicting emotions after three 1024-neuron dense layers, improving accuracy on the RAVDESS from 81.04% [28] to 87.50%. Despite a single convolution per modality, the high neuron counts of the model increased the complexity. A sparser dense layer (128 units) with additional convolution layers helped reduce computational complexity and achieved 86.00% accuracy, without the chromagram features [30].

Input size has a major impact on computational complexity and increases overhead [31]. LSTM usage in CNN models improves accuracy but escalates computation complexity [32]. This complexity limits real-time applications on resource-constrained devices. Thus, there is a growing need for compact and lightweight CNN solutions to improve AVER system robustness while meeting deployment constraints [26], [33], [34]. CNNs offer powerful tools for AVER, but in order to achieve human-level performance, complex architectures with enormous parameters are required. This poses challenges for deployment on resource-constrained devices [26], [35], [36]. Optimally stacked convolutions and hyperparameters are essential to realize competitive AVER performance while promoting a lower memory footprint and reducing power consumption [37], [38].

C. Input Features in Audio-Visual Emotion Recognition Model

The efficacy of the AVER model relies not only on facial images but also on auditory features, where feature selection is crucial [39]. Key auditory features include prosodic features (e.g., pitch, energy), qualitative features (e.g., formant), and derived features like MFCCs [40]. Prosodic features capture

emotional tone but offer limited discriminative power and are sensitive to speaking variations [41]. Qualitative features provide vocal characteristics but are computationally intensive, particularly in noisy environments [42]. MFCCs are widely adopted for their computational efficiency and superior feature representation [43], [44]. While combining features can enhance emotion recognition, it may cause overfitting and inefficiency [45], [46]. Feature selection must therefore be optimized for accuracy and complexity, especially in real-time applications [47].

MFCCs can model speech dynamics effectively by integrating time-domain and frequency-domain information [48]. The robustness to signal variations is being supported by Discrete Cosine Transform (DCT), which isolates vocal tract features from noise, and cepstral mean subtraction, which reduces static channel noise. The MFCCs provide compact, meaningful representations that simplify analysis while preserving essential speech characteristics [49]. Their alignment with human auditory perception via logarithmic scaling is found to enhance emotion recognition performance. Additionally, the MFCCs handle distortions from speaker-microphone distance effectively, making them ideal for robust emotion recognition [50].

III. Methodology

The following subsections detail the components of the customized CNN architecture for the spatial and temporal AVER model. The following subsection "Spatial Modelling with Emotional Faces" explains about facial image processing, while "Temporal Feature Extraction from Audio Speech Samples" focuses on auditory analysis; "Spatio-Temporal Dynamics Interplay" explores the integration of spatial and temporal dynamics, while "Model Optimization Strategy" and "Benchmark Datasets" discuss optimization techniques and evaluation datasets. Nomenclatures are summarized in Table I.

A. Spatial Modelling With Emotional Faces

Frames were extracted from the input video sequence at one-sixth intervals, resulting in six images per clip, using Pydub (version: 0.25.1) in Python [51], guided by [30]. Each actor's face was detected and extracted using the Haar cascade face detection algorithm to remove background noise [52], then converted into grayscale. Grayscale conversion reduced overall model complexity by at least threefold, enhancing focus on relevant facial expressions. The transformed images were resized to 64 × 64 to match the input size of the proposed facial sub-network. The preprocessing pipeline is described by Fig. 1. The proposed facial sub-network, a 3-layer 2D CNN, aimed to extract emotion-related spatial components from the 64 × 64 facial images. Initial convolution with a 3 × 3 receptive field and stride of 1, focused on local spatial features, as per findings in facial emotion recognition (FER) [53] and [54]. The convolution produced 16 distinct feature maps describing low-level facial representations. The 2D convolution operation of a single kernel is generally expressed as:

$$y_{i,j} = \sum_{m=0}^{K_h-1} \sum_{n=0}^{K_w-1} X_{i+m,j+n} \cdot W_{m,n} \tag{1}$$

TABLE I
List of Nomenclatures

Nomenclature	Description
$y_{i,j}$	Output activation value of row *i* and column *j*
K_h	Height of kernel
K_w	Width of kernel
$X_{i+m,\ j+n}$	Input activation value within the receptive field of kernel
$W_{m,n}$	Kernel coefficients
Y	A complete convolution operation
C_i	Channel of input feature map
y_k	Input feature map that contains row *i* and column *j*
b	Bias term
Y_h	Height of input feature map
Y_w	Width of input feature map
C_o	Channel of output feature map
θ_{conv}	Learnable parameter of convolution layer
Z	Output after activation function
s_o	Output spatial size that contains height and width
S	Stride size
s_i	Input spatial size that contains height and width
P	Padding size
K	Kernel size that is equivalent to K_h or K_w
X_k	Quantized MFCC
k	A set of numbers representing MFCC number
N	Number of MFCC
x_n	Mel log powers from Mel filter bank on Mel scale
b	Segment size that is equivalent to sampling time(s) × sampling rate / hop length
m	Segment averaged MFCCs
M	MFCCs 1D vector
Z_{concat}	Concatenated tensor
$Z_{spatial}$	Flattened embedding of spatial model
$Z_{temporal}$	Flattened embedding of temporal model
φ	Non-linear activation values of dense layer
Z_w	Width of Z_{concat}
W_n	Weights of the dense layer
b_n	Bias term of dense layer
n	Number of neurons
σ	Probability distribution
$e^{\varphi 2(i)}$	Exponential of activation value of second dense layer
L	Loss function
γ	One-hot encoded label
C	Number of classes

where $y_{i,\ j}$ is the output activation value of row *i* and column *j*, K_h and K_w are the height and width of applied kernel, $X_{i+m,\ j+n}$ is the input activation value within the receptive field of kernel, $W_{m,n}$ is the kernel coefficients. Then, a complete convolution operation producing an output Y from a single kernel, was executed for a total of number equivalent to input channel, C_i. The final expression can be written as:

$$Y = \sum_{k}^{C_i} y_k + b \tag{2}$$

where Y is the final output feature map for a single kernel, y_k is each input feature map that consists of row *i* and column *j*, and b is the bias term. The measure of computation complexity, hereinafter referred to as number of floating-point operations (FLOPs) for a full convolution can be obtained via:

$$FLOPs_{conv} = (2 \times Y_h Y_w C_i C_o K_h K_w) + Y_h Y_w C_o \tag{3}$$

where the terms Y_h, Y_w, and C_i are the height, width, and channel of input feature map, respectively, and K_h, K_w, and C_o are the height, width and channel of output feature map. The learnable

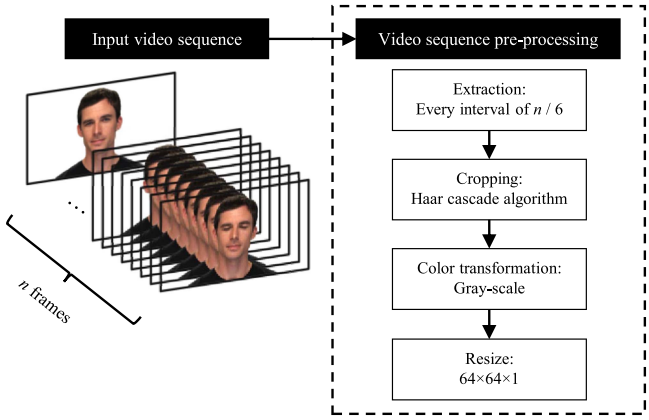

Fig. 1. Preprocessing pipeline of spatial model input. Each input video sequence will produce 6 facial images as the input for proposed spatial model.

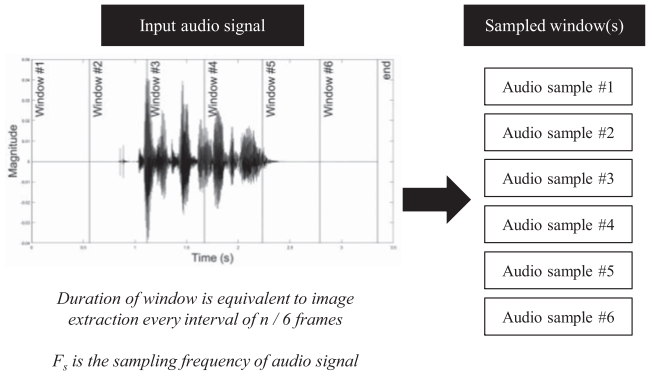

Fig. 2. Segmentation pipeline of temporal model input. Each sampled window will produce 40 MFCCs, resulting in each input audio signal producing 40 × 6 MFCCs.

parameter size of convolution, θ_{conv} is simply the kernel volume added with size of the bias term, given as:

$$\theta_{conv} = C_i\ C_o K_h K_w + C_o \tag{4}$$

Then, the corresponding feature map was further convolved with two similar 3 × 3 kernel and stride size that would produce 32 and 64 feature maps, respectively. Each of the convolutions was paired with ReLU to introduce non-linearity to the model through the piecewise linear behavior of the function and the final output from the convolution, Z is given by:

$$Z = ReLU\ (Y) = \begin{cases} Y & \text{if } Y > 0, \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $Y \in \{Y_{1,11}, Y_{1,12}, \ldots, Y_{i,j,Cout}\}$. In fact, increasing the channel of a single convolution operation could have a higher tendency to seek after finer and abstract details over a specific, extracted facial feature. However, it simultaneously escalates the complexity of optimization on each kernel, counteracting on the benefit of higher representation power from the convolution [55]. Hence, a dropout layer was added after the last convolution to enforce the learning of useful embeddings and prevent overfitting, as adopted in [56]. To further suppress any redundance in local regions, the spatial resolution between each successive convolution was down sampled into half of its original size through max pooling of size and stride of two. Indeed, while max pooling could potentially discard fine-grained image details, it is also likely to discard artefacts in macro-expression analysis, thus rendering the possibility of capturing more differences between emotions at a global level. Moreover, the FLOPs of subsequent convolution layer after the max pooling operation could be further reduced since the spatial size (i.e., Y_h and Y_w) of input feature map of the subsequent convolution layer is directly affected by the output height and width, s_o from max pooling. Assuming identical spatial resolution, the output spatial size, s_o of max pooling can be computed through:

$$s_o = (S + s_i + 2P - K)\ /S \tag{6}$$

where s_i is the spatial size (i.e., Y_h and Y_w) of input feature map, P is the padding size, K is the kernel size and S is the stride size. Equation (6) can be similarly used to calculate spatial size of convolution, Y_h and Y_w with s_i referring to the output spatial size from max pooling layer. Given our convolution configuration, $K = 3$, $P = 0$ and $S = 1$, the Y_h and Y_w are essentially $(s_i - 4)$ / 2 smaller than when without max pooling. The computation incurs null FLOPs to the model.

B. Temporal Feature Extraction From Audio Speech Samples

To effectively extract the synchronized temporal information for the spatial modelling above, the speech audio samples were also segmented by every one-sixth interval over the whole duration of the signal without overlapping between the intervals. The detail of the segmentation pipeline is described by Fig. 2. While overlapping segmentation could help to maintain the context and continuity, at the same time, it introduces redundancy and semantic affinity between the adjacent segments that might disrupt the spectral peculiarity of a short duration speech sample [57]. The segmented speech audio signal was then quantized into cepstral representations through MFCC that has been proven to be a robust descriptor for speech emotion recognition (SER) [57], [58]. The current study extracted 40 MFCCs from each segmented sample rather than the common 13 MFCCs to extract more fine-grained high frequency spectral envelopes that are beyond the fundamental frequency information such as pitch. The quantization of MFCC from Discrete Cosine Transform (DCT), specifically type II as detailed in [59], [60], is expressed as:

$$X_k = 2 \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \tag{7}$$

where $k \in \{0, 1, 2, \ldots, N-1\}$, N is the number of cepstral coefficients, and x_n is the Mel log powers from Mel filter bank on Mel scale. The computation was done via librosa (version: 0.10.1) in Python [61] using the default values and the sampling rate was from the speech audio source. The output from the function was a matrix of size $k \times b$ where b is the Mel spectrogram of size sampling time(s) × sampling rate / hop length and the default value for hop length was 512. The resulting MFCCs, m was transposed and averaged to obtain a 1D vector, denoted as $M \in \{m_1; m_2; m_3; \ldots; m_{40}\}$.

The row vector M, was then used as an input for the proposed speech audio sub-network that is a 2-layer 1D CNN, primarily focusing on distinct patterns in the cepstral domain. The selection of input type between raw signal and MFCC vector was application specific. Using MFCC as an input source provides advantages such as: 1) low-dimensional compact feature contains less redundant information, 2) focus on relevant spectral information and robust against noise, 3) capable of capturing prosodic information (i.e., pitch, rhythm, intonation). Overall, MFCC should be able provide a more instinctive representation to the model and reduce the learning complexity. In fact, there show no superiority over one another (i.e., raw audio signal versus MFCC representation) in SER tasks using CNN model, despite MFCC capabilities in raw speech audio signal feature extraction [62]. The network design was inspired by [63], where intriguing SER performance was achieved with only two convolutions. To expand its capabilities without increasing the complexity, we introduce progressive channel expansion instead of the fixed 128 channel output from each convolution. This favors lower learnable parameters and computation complexity, according to (3) and (4). The first convolution utilized a 5×1 kernel and stride of 1 to capture the cepstral dynamics between each five bands of frequencies and produce another 16 sets of specified cepstral dynamic pattern. In this case, 1D convolution is used and its mathematical expression is alike (1), except, the width of the kernel, K_w is set to 1. Then, these 16 sets of patterns were convolved again with similar kernels and stride to further identify the key changes in spectral semantics and ultimately resulted in 32 sets of features that were based on the dynamics of cepstral coefficients. A similar approach for activation function and max pooling from the facial sub-network above is adopted here. Dropout was also added for the second convolution to ensure non-repetitive learning on the dynamics.

C. Spatio-Temporal Dynamics Interplay

Before the concatenation, both the tensors of spatial and temporal embeddings were flattened to match the tensor dimension and synchronize the spatio-temporal information. We preferred flattening over global pooling methods to avoid information loss due to arithmetic operations [64]. The concatenated tensor is denoted as $Z_{concat} \in \{Z_{spatial}, Z_{temporal}\}$ where $Z_{spatial}$ and $Z_{temporal}$ contains all the embeddings flattened from spatial model and temporal model, respectively. The resulting concatenated tensors Z_{concat} are then fully mapped to a dense layer (i.e., fully connected layer) with 16 neurons for RAVDESS and MEAD model and 14 neurons for SAVEE, to define the relationship between the spatial and temporal tensors, given by the non-linear transformation equation:

$$\varphi_1\ (n) = Z_{concat}\ \cdot W_n\left(\varphi_1\right) + b_n \tag{8}$$

where φ_1 is the non-linear activation values of first dense layer, W_n is the weights of the dense layer, b_n is the bias term, and n is the number of neurons. The computation complexity and size of learnable parameter is alike to (3) and (4), except here, $C_o = n$ and $C_i = Z_w$ where Z_w is the width of Z_{concat}, with the rest all set to 1. Dropout was added before the last dense layer for similar intention. We then allocated 8 neurons for RAVDESS and MEAD model and 7 neurons for SAVEE in the last dense layer, according to their total emotion classes. The mathematical expression is defined as:

$$\varphi_2\ (n) = \varphi_1\ \cdot W_n\left(\varphi_2\right) + b_n \tag{9}$$

where n corresponds to the number of emotion classes. Following this is the probability distribution of the final dense layer, obtained using the Softmax function, which is denoted as:

$$\sigma\ \left(\varphi_2\left(i\right)\right) = e^{\varphi_2(i)} / \sum_{j=1}^{n} e^{\varphi_2(j)} \tag{10}$$

where σ is the probability distribution, $e^{\varphi 2(i)}$ is the exponential of activation value of i-th element, and n is the neurons of φ_2. Details of the proposed spatio-temporal model are illustrated in Fig. 3. Convolution kernel size, kernel count, max pooling size, and dense layer neurons are specified for each layer. The output feature map dimensions—height (H), width (W), and channel (C)—are also provided for each layer.

D. Model Optimization Strategy

The proposed model was trained with an Adam optimizer, with a default learning rate of 0.001, exponential decay rate for first moment estimates of 0.9, second moment estimates of 0.999, and ε of 10^{-7}. The Adam optimizer was chosen for its faster convergence and superior optimization of sparse gradients via exponential decay [65], [66]. Categorical cross-entropy was used for backpropagation, expressed as:

$$L\ (\gamma, \sigma) = -\sum_{i=1}^{C} \gamma_i \cdot \log \sigma_i \tag{11}$$

where γ is the one-hot encoded label for the observed data, C is the number of classes, and σ is the calculated probability from (10). The fully expanded partial derivative of (11) with respect to φ_2 is derived as:

$$\delta L / \delta \varphi_2 = \sigma - \gamma \tag{12}$$

The model was trained with a batch size of 128 and early stopping was adopted to prevent model overfitting from prolonged training epochs. The early stopping criteria were based on the validation loss, i.e., when the $\Delta loss$ is smaller than 0.001 for 30 consecutive epochs, the training will be terminated and the model's weight of 30 epochs earlier will be restored.

E. Benchmark Datasets

Three popular datasets for emotion recognition are acquired to assess the performance of the proposed model: 1) SAVEE, 2) RAVDESS, 3) MEAD. Each of the dataset contains video clips of an average of 3 seconds time and the details are described as follows. The trained model was evaluated using the confusion matrix accuracy, precision, recall, and f1-score. Model complexity was quantified by the number of parameters, FLOPs, memory requirements in FP32, arithmetic intensity (FLOPs/memory), and average inference speed for 10 CPU forward propagations.

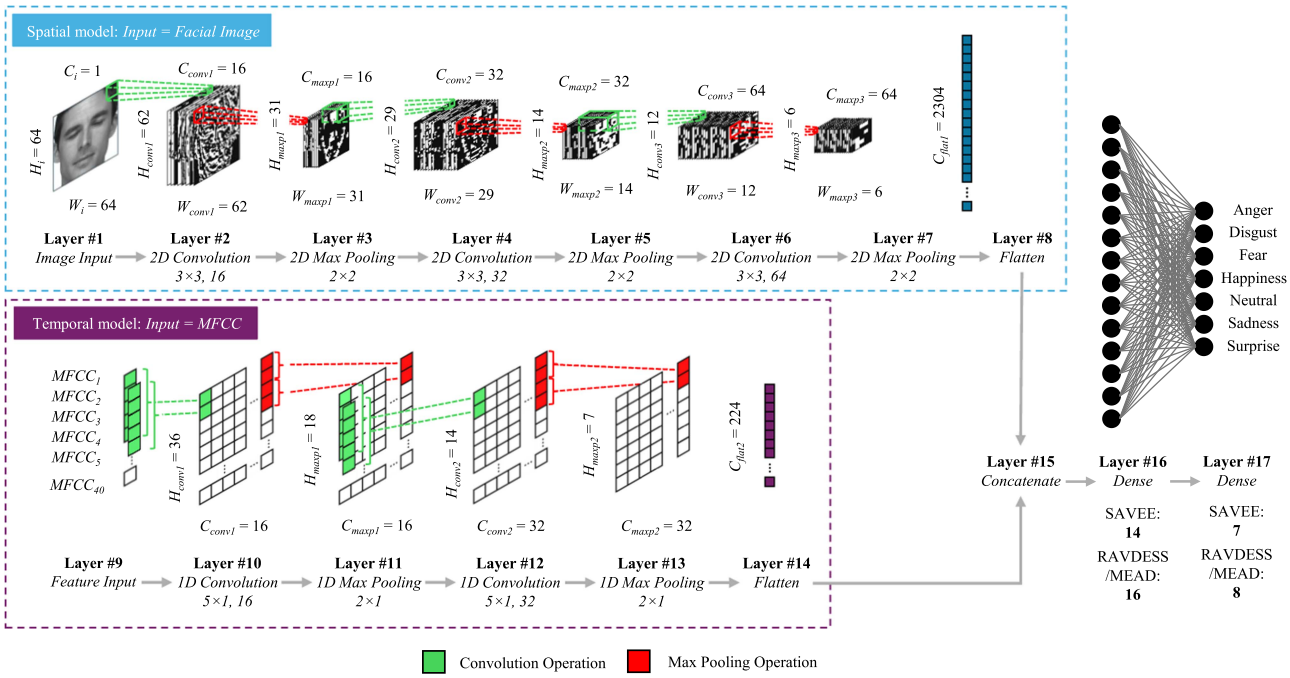

Fig. 3. Proposed spatio-temporal model. Dropout layers are added before Layer #8 and Layer #14 in respective spatial and temporal model. The neurons of Layer #16 are defined according to datasets, where SAVEE=14, and RAVDESS/MEAD=16. Similar for Layer #17, SAVEE=7, and RAVDESS/MEAD=8. Another dropout layer is added before Layer #17. The ratio of all dropout layers is 0.3, 0.4, and 0.5, for MEAD, RAVDESS, SAVEE, respectively.

1) SAVEE: This dataset comprises recordings from four male actors (ages 27–31) from the University of Surrey. Each actor recorded 120 utterances, totaling 480 utterances. Each recording includes 15 TIMIT sentences per emotion: 3 common, 2 emotion-specific, and 10 generics. TIMIT, designed by Texas Instruments, Inc., and the Massachusetts Institute of Technology [67], provides phonetically balanced transcriptions. The dataset includes seven emotion categories: "Anger", "Disgust", "Fear", "Happiness", "Sadness", "Surprise", and "Neutral". Videos were recorded at 60 fps in .avi format, and audio at 44.1 kHz in .wav format. Each emotion class contains 60 audio-visual files, except "Neutral", which has 120. Video clips are segmented into six parts, yielding 2880 facial images and audio segments.

2) RAVDESS: This dataset features 24 professional actors, balanced by gender, vocalizing two lexically matched statements in a native North American accent. Each emotion is expressed at two intensity levels, with an additional neutral expression. Eight emotions are included: "Anger", "Disgust", "Fear", "Happiness", "Sadness", "Surprise", "Neutral", and "Calm", resulting in 1440 audio-visual clips. Videos are recorded in 720p (H.264, .mp4), and audio at 48 kHz (.wav). The dataset includes 8640 facial images and audio segments. To address limited lexical variability, Gaussian noise (variance 0.01) was applied to facial images, and a time stretch of 0.8 was applied to audio segments.

3) MEAD: This dataset comprises 60 gender-balanced actors, aged 20–35, fluent in English, with some having acting experience. Video clips were recorded from multiple angles in a controlled environment to minimize noise distortion, though only frontal views are used in this study. The dataset includes eight emotions ("Anger", "Disgust", "Fear", "Happiness", "Sadness", "Surprise", "Neutral", and "Contempt") at three intensity levels, except for "Neutral". Actors performed 30 sentences per emotion and 40 for "Neutral". A total of 30412 video clips were recorded in 1080p, 30 fps (.mp4), and audio in 48 kHz (.m4a). The dataset uses a structured data collection approach and guidance team led by a professional actor, ensuring natural and consistent expressions. While such a structured approach could be beneficial under certain circumstances (e.g., talking-face generation tasks), it does not ensure sufficient variability between actors, potentially leading to limited features to be captured for each emotion class.

IV. Result

A. Implementation Configuration

The hardware for model training is a mobile machine with 12th Gen Intel Core^{TM} i5-12500H CPU (~2.50 GHz), NVIDIA GeForce RTX3060 Mobile GPU (6 GB VRAM) with 3840 CUDA cores, and 32 GB of dual-channel DDR5 4800 MHz RAM. Python (version: 3.9.12) with Anaconda distribution (version: 4.12.0) was implemented for the coding task execution. Image pre-processing steps were implemented using Python package *OpenCV* (version: 4.6.0.66) [68]. All the implemented datasets followed the 80/20 train test stratified split rule, then the training set was further divided by 80/20 based on train validation stratified split rule. This stratified data split strategy ensures the reported results are not overestimated or underestimated for large datasets such as the RAVDESS and MEAD, as discussed in [69].

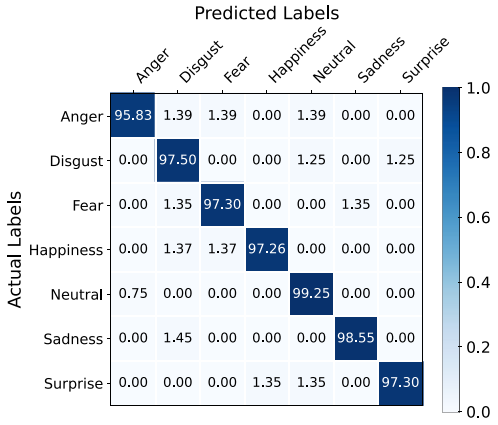

Fig. 4. Confusion matrix of proposed model for SAVEE dataset.

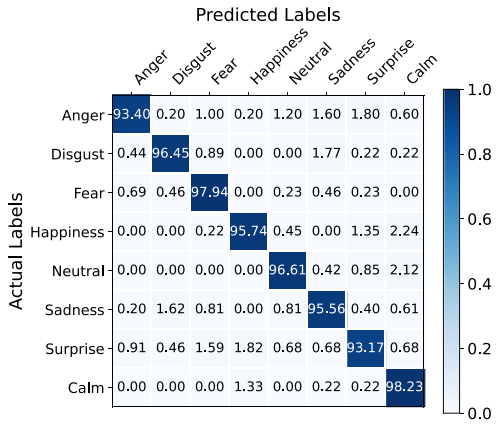

Fig. 5. Confusion matrix of proposed model for RAVDESS dataset.

B. Emotion Classification With SAVEE and RAVDESS

The performance of the proposed spatio-temporal classifier for SAVEE and RAVDESS is shown in Figs. 4 and 5, with lower accuracy observed for "Anger" in SAVEE and "Surprise" in RAVDESS. In SAVEE, "Anger" was often misclassified as "Disgust", "Fear", or "Neutral", while in RAVDESS, it was misclassified across all emotions. Similarly, "Surprise" was frequently confused with "Happiness" and "Neutral", reflecting overlapping features. Prior studies [70], [71] attribute misclassification between "Anger" and "Disgust" to shared facial cues like furrowed brows, narrowed eyes, and raised upper lips [72]–[74]. Overlapping acoustic features, such as pitch, also contribute to confusion between "Anger" and "Surprise" [75], [76], with shared visual traits like a raised upper lip further compounding errors [77]. These challenges underscore the model's limitations in distinguishing subtle emotional nuances, preventing perfect accuracy across all categories.

To evaluate the proposed model on a larger dataset, the MEAD dataset [19] was utilized, marking its first application in general AVER, as it is typically used for reconstructing facial motions from speech input. The model achieved an average accuracy of 98.57%, with the confusion matrix shown in Fig. 6. Misclassification of "Anger" in MEAD mirrored patterns in SAVEE

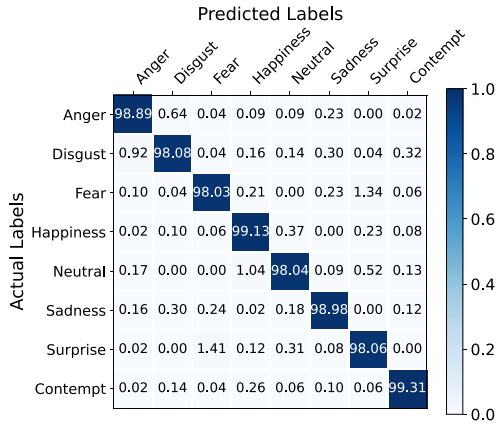

Fig. 6. Confusion matrix of proposed model for MEAD dataset.

TABLE II
LOOCV RESULT OF SAVEE DATASET

Methods	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed	SAVEE	97.57	97.60	97.57	97.59
LOOCV		97.50	97.24	97.14	97.19

and RAVDESS, except that none were predicted as "Surprise". Similarly, "Surprise" misclassification in MEAD did not include "Disgust". These results reaffirm the model's effectiveness, with consistent patterns of lower performance in recognizing "Anger" and "Surprise" observed across all datasets.

C. Validation of Model Generalizability

Leave-One-Out Cross-Validation (LOOCV) was applied to the SAVEE dataset to validate the proposed model's performance using a normal train-test-validation split. This approach ensures that reported results are not overestimated due to data leakage (e.g., the same actor appearing across emotions during training) [78]. In LOOCV, each actor is used once as the test set while the remaining actors are used for training, iterating until all four actors serve as test sets. Results are summarized in Table II. LOOCV yielded minimal differences from the proposed data split method, with decreases of 0.07%, 0.36%, 0.43%, and 0.40% in accuracy, precision, recall, and F1-score, respectively. This demonstrates the robustness and generalizability of the proposed data split method for emotion recognition across individuals.

To further evaluate generalizability, cross-dataset training and testing were conducted. The model was trained on SAVEE dataset using the optimization strategy shown in Fig. 3 and tested on RAVDESS and MEAD datasets. To maintain consistency in emotion classes, "Calm" (RAVDESS) and "Contempt" (MEAD) were excluded. This process was repeated for all datasets as training sets individually, with results detailed in Table III. Performance metrics showed varied declines (e.g., a ∼10% accuracy drop when SAVEE was used for training and RAVDESS for testing), though results remain promising and could improve with transfer learning. The model effectively

TABLE III
PERFORMANCE OF MODEL USING VARIOUS TESTING DATA

Training Set	Testing Set	Acc. (%)	Prec. (%)	Rec. (%)	F1-Score (%)
SAVEE	*RAVDESS*	84.92	85.13	84.83	84.98
	MEAD	90.61	90.39	90.43	90.41
RAVDESS	*SAVEE*	87.74	87.09	86.41	86.75
	MEAD	94.16	94.09	94.04	94.06
MEAD	*SAVEE*	84.03	83.17	82.56	82.86
	RAVDESS	67.11	67.30	66.60	66.95

extracts common emotion features when trained on SAVEE or RAVDESS, datasets designed for emotion recognition. However, performance declines when trained on MEAD, a dataset for facial motion reconstruction, achieving 84.03% accuracy on SAVEE (13.54% lower than proposed) and 67.11% on RAVDESS (28.78% lower than proposed). The performance drop when testing the MEAD-trained model on RAVDESS compared to SAVEE is likely due to differences in diversity, including expression styles, accents, and emotional intensity. SAVEE, with fewer actors, has less variability, making it easier for the model to adapt. RAVDESS uses an independent approach where actors freely interpret emotions, resulting in greater variability in intensity and clarity, which complicates generalization. On the other hand, MEAD ensures controlled emotional expressions through continuous professional guidance, leading to more consistent portrayals. This makes the MEAD model perform better on SAVEE, but it struggles with the greater variability in RAVDESS, resulting in weaker performance.

V. DISCUSSION

A. Comparison With State-Of-The-Arts in SAVEE and RAVDESS

The contemporary state-of-the-arts in AVER with model-level fusion using SAVEE and RAVDESS dataset (dated from the year 2021 to 2024) are tabulated in Table IV. Authors of [30] proposed a multimodal classifier of 3-layer 1D CNN for speech audio cues and 8-layer 2D CNN with LSTM for facial images for the RAVDESS dataset, and 2-layer 1D CNN plus 6-layer 2D CNN with LSTM for the SAVEE dataset, respectively. The study experimented with multiple combinations of speech audio and facial feature extractors, with the best combination achieving an accuracy of 99% for ER task on SAVEE and 86% on RAVDESS. Such discrepancy in classification performance for the two datasets suggests that 1) SAVEE could be a relatively easy dataset, since it was only recorded by four male actors with much phonetically balanced sentences, while RAVDESS is a gender-balanced dataset but with extremely limited lexical, leading to potentially higher variations in emotion expression but relatively constant in verbal expression, 2) using LSTM modules might not be an ideal solution when there is limited in between-group variation (i.e., limited lexical between emotions). Nevertheless, the implementation of LSTM also significantly increases the demand of computation resources without notable improvement in accuracy [79], [80]. Authors of [81] also reported high accuracy on SAVEE dataset and performed less well on the RAVDESS dataset with the proposed AVER model, scoring an accuracy of 94.99%. The model performed the worst for "Sadness" emotion, only able to predict at a coin-flipping probability (i.e., an accuracy of 57%). Similar performance on "Sadness" emotion was observed in [30], where the lowest accuracy was 53%. The performance on SAVEE worsened by 3.52% when the audio model in [81] was replaced with a 3D CNN, despite an increment of 67% the number of parameters [82]. This indicates that simply increasing the complexity of the speech audio model and types of auditory features, specifically for SAVEE dataset, does not necessarily benefit the performance. In fact, it is a sign of overparameterizing the model. Instead of using LSTM to learn the dependencies on the chronological sequence between the frames, [83] proposed an approach with 3-layer 2D CNN for both speech audio and facial representation, that is duplicated 12 times to accommodate the randomly selected temporal segments over the course of recording duration. The model yielded an averaged accuracy of 78.75% in RAVDESS. Apart from the accuracy, the study demonstrated its robustness against varying temporal lengths (i.e., total recording time) and inference time that is more than sufficient for face-to-face human ER through communication, as suggested in [84]. In addition, the study also fared poorly in classifying "Sadness" (i.e., 46.88% of accuracy) emotion. Even human perception encounters difficulties in accurately identifying the "Sadness" emotion [18], [85].

Using Occam's razor principle, it is unnecessary to have complex models to be successful in AVER using speech audio and facial images. While it is convenient to adopt models that are developed for general applications using ImageNet, such as the inception modules in [81] and [82], or VGG16 variants in [86], it tends to overly invest in the number of model parameters for high accuracy. A simple ER-specialized structure with good optimization can achieve similar results [30]. We present the model's requirement on computation in Table V. The proposed model in this work achieved comparable accuracy of 97.57% in SAVEE dataset and 95.89% in RAVDESS dataset, with a significant reduction in model complexity. Specifically, it is 99.88% fewer parameters and FLOPs compared to the best performer in SAVEE, proposed by Sarafi et al. [81], along with a 99.26% reduction in memory requirements; it also exhibits reductions of 99.94% in parameters, 99.65% in FLOPs, and 98.39% in memory requirement compared to the best model for RAVDESS, proposed by Bilotti et al. [86], with only 0.06% drop in accuracy. The substantial reduction in model complexity of the proposed model also resulted in 75.48% (i.e., 579 ms) and 22.77% (i.e., 59 ms) faster inference speed than [81] and [86]. Furthermore, our proposed model strived in the weakest spot of study in [30], [81], and [83] in RAVDESS (i.e., 95.56% of accuracy for "Sadness" emotion), demonstrating well-generalized emotion features while retaining the performance on both datasets.

B. Ablation Study on Proposed Model

To assess the limitations and failure points of the proposed model, additional experiments were conducted on SAVEE and

TABLE IV
STATE-OF-THE-ART FACIAL-SPEECH MULTIMODAL EMOTION RECOGNITION ARCHITECTURES FOR SAVEE AND RAVDESS DATASET

Author (year)	CNN Architecture		Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
	Speech audio	*Facial*					
Middya et al. (2022) [30]	1D (mix of auditory features)	2D-LSTM	SAVEE	99.00	99.86%	99.86%	99.86%
Sharafi et al. (2022) [81]	1D (mix of auditory features)	DSN-DTN-Bi-LSTM		99.75	99.58%	99.57%	99.58%
Sharafi et al. (2023) [82]	3D (mix of auditory features)	DSN-DTN-Bi-LSTM		95.48	95.12%	94.71%	94.91%
Proposed	**1D (MFCC)**	**2D**		**97.57**	**97.60%**	**97.57%**	**97.59%**
Radoi et al. (2021) [83]	2D (log Mel spectrogram)	2D	RAVDESS	78.75	78.80%	78.91%	78.85%
Middya et al. (2022) [30]	1D (mix of auditory features)	2D-LSTM		86.00	85.93%	85.63%	85.78%
Sharafi et al. (2022) [81]	1D (mix of auditory features)	DSN-DTN-Bi-LSTM		94.99	96.55%	95.75%	96.15%
Bilotti et al. (2024) [86]	2D (Mel spectrogram)	2D		95.95	*N/A*	*N/A*	*N/A*
Proposed	**1D (MFCC)**	**2D**		**95.89**	**95.91%**	**95.89%**	**95.90%**

DSN = deep spatial network; DTN = deep temporal network

TABLE V
COMPUTATION REQUIREMENTS FOR STATE-OF-THE-ARTS

Author (year)	Dataset	Param (M)	FLOPs (G)	Mem. (MB)	Inf. time (ms)
Middya et al. (2022) [30]	SAVEE	5.61	1.178	8.58	277
Sharafi et al. (2022) [81]		50.80	12.307	126.98	767
Sharafi et al. (2023) [82]		64.39	289.526	1631.41	3100
Proposed		**0.06**	**0.014**	**0.93**	**188**
Radoi et al. (2021) [83]	RAVDESS	1.63	0.042	4.66	246
Middya et al. (2022) [30]		8.60	1.797	9.46	278
Sharafi et al. (2022) [81]		50.80	12.307	126.98	767
Bilotti et al. (2024) [86]		108.00	4.105	58.04	259
Proposed		**0.06**	**0.014**	**0.93**	**200**

Param = number of parameters; FLOPs = floating point operations; Mem = memory; Inf. Time = Inference time

RAVDESS datasets using the same optimization strategy. The ablation study focused on the impact of modalities (spatial and temporal), convolution kernel size of 2 up to 12, and dropout rates of 0 to 0.9. Occlusion tests on facial images (eyes, eyebrows, and mouth) and auditory MFCC features (lower- and higher-order) were performed to evaluate the significance of these features in the model's predictions. Additionally, Integrated Gradient (IG) [87] was applied for validation. The negative gradient does not indicate redundancy but reflects a negative contribution to the prediction score. The details of the ablation result can be found in Supplementary material, and we summarize the findings below.

1) Importance of Spatial and Temporal Models in AVER: The ablation result on the impact of modalities shows dominance of facial features in emotion recognition in general. However, the model achieved better generalization and performance after integrating with auditory features, with a subtle increase in cost of model complexity. The resulting IG revealed a higher dependency of model on facial features when trained with limited-actor datasets like SAVEE while both facial and auditory features are critical for actor-diverse datasets like RAVDESS.

2) Effects of Convolution Kernel Sizes: For convolution kernel size ablation, the obtained results shows that larger kernel size does not necessarily improve performance as they may also capture irrelevant features that leads to poorer performance, especially when the model is trained with limited-actor datasets like SAVEE.

3) Effects of Dropout Rate: For dropout rates, there exists a fine balance between feature retention and performance, where limited-actor datasets like SAVEE requires a higher dropout rate of 0.5, while actor-diverse datasets like RAVDESS requires lower at 0.4, which is less critical for regularization.

4) Occlusion Test of Facial Features: The occlusion of facial images shows a negligible drop in average performance for both datasets when compared to spatial model trained with no occlusion. However, fluctuations are observed in individual emotion classes. The general visual inspection revealed distinct expression patterns between both datasets for "Happiness" and is likely due to the preferences of expression cues for social interaction in different ethnicity. Despite performance variations, no facial feature dominated across emotion classes, underscoring the relevance of all features in emotion recognition, consistent with findings in [88]. The IG of facial images of both datasets in Fig. 7 shows the occluded regions in the facial images are similarly focused for emotion recognition, with added attention to cheek area for RAVDESS. We also found that RAVDESS had better gradient representation than SAVEE, reflecting the importance of actor diversity in dataset. The gradient flow around

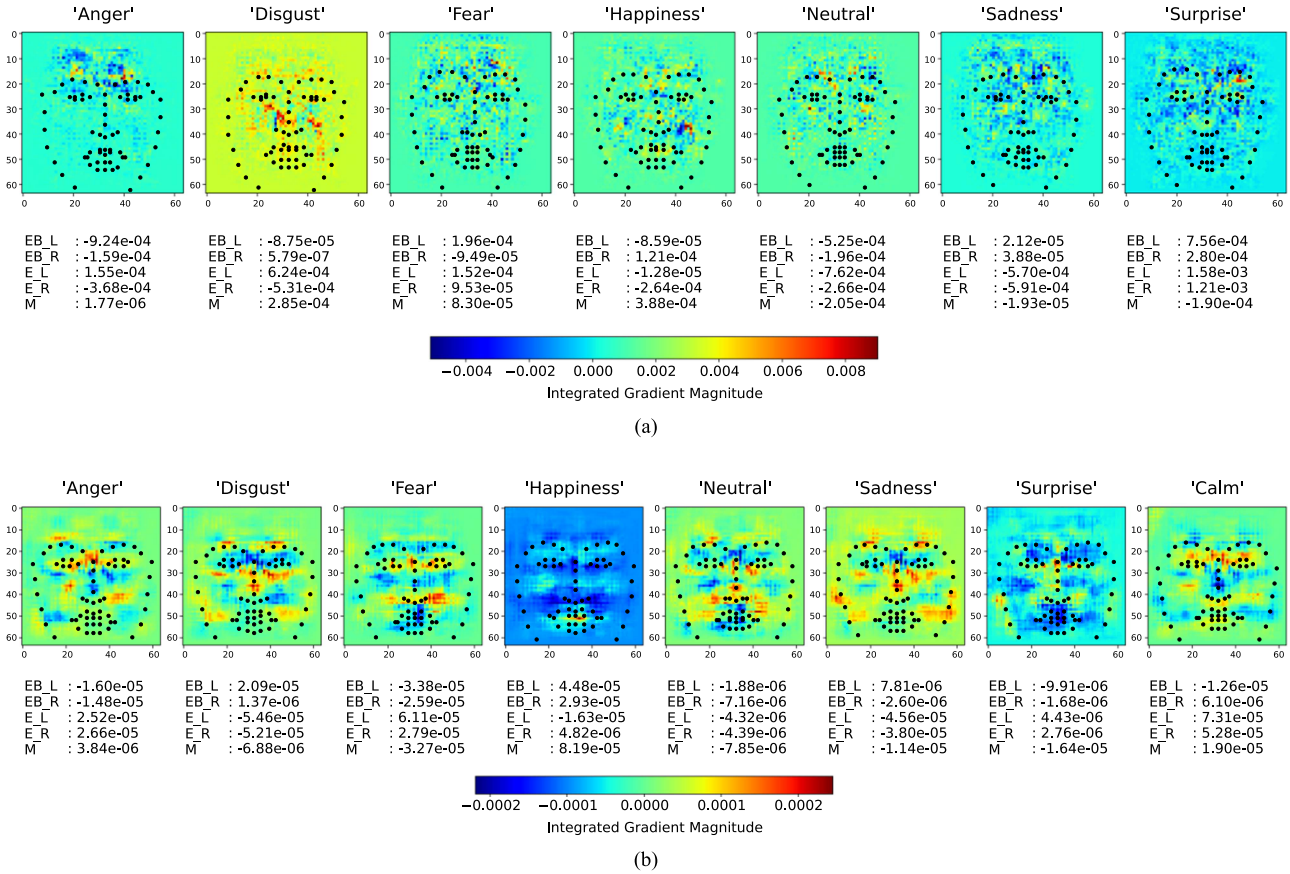

Fig. 7. Magnitude of integrated gradients based on extracted facial landmarks for each emotion classes in (a) SAVEE (b) RAVDESS. The mean magnitudes of left (L) and right (R) of EyeBrows (EB), Eyes (E), and Mouth (M) bounding box is located below the heatmap. The facial landmarks are laterally inverted. Any magnitude that is approaching 10e-06 can be considered as near zero. The color of the gradients represents the contribution levels. For instance, in SAVEE, gradients near light blue does not contribute towards prediction score, while in RAVDESS is green.

the Duchenne markers on "Happiness" of both datasets aligns with [89], confirming that they are key facial cues for perceiving happiness and is captured by the proposed spatial model.

5) Occlusion Test of Audio Features: The results from occlusion of lower- and higher-order MFCCs demonstrates almost no improvement on both datasets, suggesting that general representations from lower-order MFCCs alone are insufficient for emotion recognition, while subtle frequency details from higher-order MFCCs also fail to capture emotions independently. The evenly distributed gradients observed from the applied IG in Fig. 8 explain the dependency of the features for emotion recognition and resulted in a significant decline in performance during the occlusion test. The exhibited properties underscore the necessity of both lower- and higher-order for robust emotion recognition.

C. Target Embedded Device for Potential Deployment

For real-time ER on embedded computing boards (ECBs) like the Raspberry Pi 4B, models need to be around 0.3 MB in size [90], where none from Table IV fulfilled such a requirement. However, our proposed model, after conversion to TensorFlow Lite (TFLite), can meet these requirements with quantization-aware training, that can reduce model size by up to 75% without performance loss [91]. Additional ECB options and their theoretical performance estimates are detailed in Supplementary Table VI. Key options include Raspberry Pi and NVIDIA Jetson, known for their cost-effective AI modeling capabilities. While these devices may not be optimized for complex CNN models due to factors like low arithmetic intensity and FP32 peak performance, they can still support model implementation given the required FLOPS stay within FP32 peak performance limits. Fig. 9 illustrates implementation feasibility using the roofline model. The GFLOPS for state-of-the-arts are estimated based on inference time from Table V. Except for [82], all models can be deployed on the targeted ECBs, with expense on more arithmetic operations per data byte transfer. Although the model from [83] can be optimally deployed on two additional options from NVIDIA based on arithmetic intensity, its prediction accuracy is 17.14% lower than our proposed model with 2.43 times more GFLOPS. The surplus in GFLOPS is expected to introduce more latency, as confirmed by the testing on our machine's CPU.

D. Limitations and Future Directions

Despite the proposed model's promising complexity-to-performance ratio, several limitations remain. The efficiency was evaluated theoretically, not on ECBs. Future work will

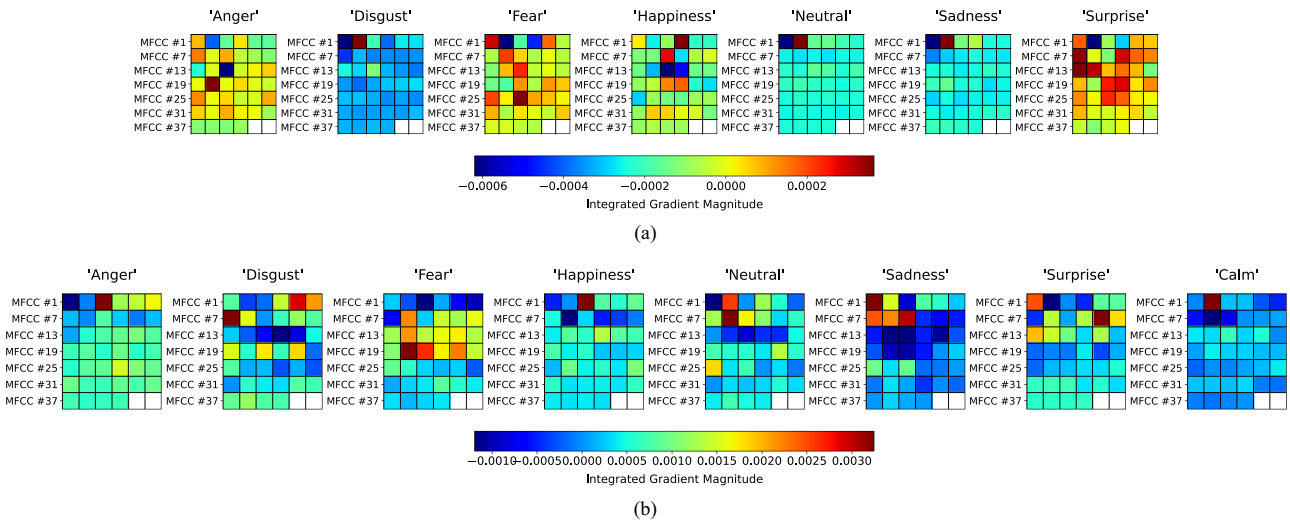

Fig. 8. Magnitude of integrated gradients of MFCCs for each emotion classes in (a) SAVEE (b) RAVDESS. The color of the gradients represents the contribution levels. For instance, in SAVEE, gradients near yellow does not contribute towards prediction score, while in RAVDESS is light blue.

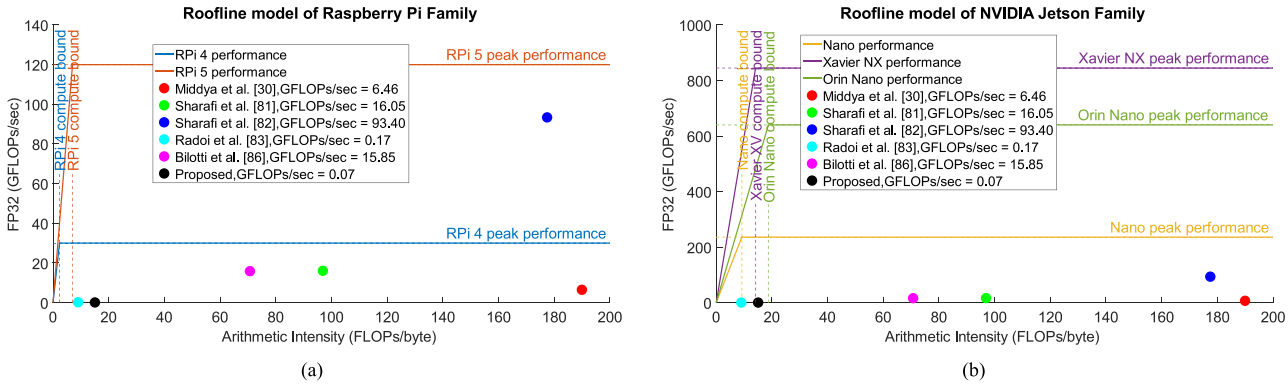

Fig. 9. Roofline model for (a) Raspberry Pi (b) NVIDIA Jetson family ECBs. Models beyond the compute bound are still deployable, with extra operation per data byte transfer. GFLOPs/sec for state-of-the-art is estimated with the inference time in Table V. Models beyond the ECB peak performance is difficult to deploy.

focus on real-time ER implementation on ECBs. Secondly, the proposed model's generalizability was tested only on individual datasets. Future work should employ different datasets with identical emotions for training and testing to better assess generalizability. Additionally, this study did not exhaustively explore convolution hyperparameters, particularly asymmetric kernel sizes. Future research could develop specialized kernels to capture specific facial features and fully automate ER tasks. The model's ability to extract features for closely related emotions can be enhanced by incorporating facial landmarks detected through advanced algorithms. The benchmark datasets, particularly SAVEE, lack actor diversity, as indicated by gradients in the ablation study. Increasing the number of actors within a dataset would improve gradient flow and enhance recognition of emotions like "Happiness". Targeted augmentations on specific facial features, rather than entire images, are recommended to generalize the model for feature variations effectively.

SAVEE's limitations extend to its lack of diversity in gender, cultural, and linguistic backgrounds, restricting model generalizability. The dataset's small size, limited utterances, and constrained emotion intensity further reduce variability within emotion classes, making it difficult to distinguish closely related emotions with overlapping facial and vocal traits. These factors may introduce biases and limit the model's applicability to broader populations or real-world scenarios.

Future research should prioritize datasets with actors from diverse demographics, including balanced gender representation and varied cultural and linguistic contexts. Expanding recordings to include more utterances and a broader range of emotion intensities would likely enhance variability, hence supporting the development of more robust and generalizable models.

VI. Conclusion

This paper presents a multimodal CNN architecture for AVER, utilizing MFCC for speech audio and facial images as inputs. The model achieves high accuracy — 97.57% on SAVEE, 95.89% on RAVDESS, and 98.57% on MEAD — while

significantly reducing the number of parameters, thereby enhancing feasibility for real-time applications on resource-constrained devices. The proposed spatio-temporal model, combining a 2-layer 1D CNN for audio analysis and a 3-layer 2D CNN for visual analysis with model-level fusion, is significantly lightweight, offering a 99.94% reduction in parameters compared to state-of-the-arts. This efficiency enables deployment on platforms such as the NVIDIA Jetson Orin Nano without performance degradation, paving the way for broader applications of AVER in real-world scenarios.

REFERENCES

[1] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion recognition for human-robot interaction: Recent advances and future perspectives," *Front. Robot. AI*, vol. 7, Dec. 2020, Art. no. 532279.

[2] V. Savchenko, L. V. Savchenko, and I. Makarov, "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2132–2143, Third Quarter 2022.

[3] T. Dar, A. Javed, S. Bourouis, H. S. Hussein, and H. Alshazly, "Efficient-SwishNet based system for facial emotion recognition," *IEEE Access*, vol. 10, pp. 71311–71328, 2022.

[4] J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, Sep. 2023.

[5] T. Han et al., "Text emotion recognition based on XLNet-BiGRU-Att," *Electronic*, vol. 12, no. 12, pp. 2704–2704, Jun. 2023.

[6] S. M. Abdullah, S. Y. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using Deep Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 73–79, Apr. 2021.

[7] I. Madhavi et al., "A deep learning approach for work related stress detection from audio streams in cyber physical environments," in *Proc. IEEE Int. Conf. Emerg. Technol. Factory Automat.*, Vienna, Austria, 2020, pp. 929–936.

[8] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 8093–8104.

[9] T. Oyedare, V. K. Shah, D. J. Jakubisin, and J. H. Reed, "Keep it simple: CNN model complexity studies for interference classification tasks," in *Proc. IEEE Conf. Comput. Commun. Workshops*, Hoboken, NJ, USA, 2023, pp. 1–6.

[10] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[11] D. S. Maura, T. Goel, K. Goswami, D. S. Banerjee, and S. Das, "Variation aware power management for GPU memories," *Microprocessor Microsyst.*, vol. 96, no. 1, Feb. 2023, Art. no. 104711.

[12] S. Kiranyaz et al., "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, Apr. 2021, Art. no. 107398.

[13] D. Kim, H. Cho, H. Shin, S.-C. Lim, and W. Hwang, "An efficient three-dimensional convolutional neural network for inferring physical interaction force from video," *Sensors*, vol. 19, no. 16, Aug. 2019, Art. no. 3579.

[14] A. Dhoot, N. Hadj-Alouane, and M. Alouane, "2D CNN vs 3D CNN: An empirical study on deep learning-based facial emotion recognition," in *Proc. Int. Conf. Model. Simul. Intell. Comput.*, Dubai, UAE, 2024, pp. 138–143.

[15] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Mar. 2020.

[16] H. M. Shahzad, S. M. Bhatti, A. Jaffar, M. Rashid, and S. Akram, "Multimodal CNN features fusion for emotion recognition: A modified xception model," *IEEE Access*, vol. 11, pp. 94281–94289, 2023.

[17] S. Haq and P. Jackson, "Multimodal emotion recognition," in *Machine Audition: Principles, Algorithms and Systems*, Hershey, PA, USA: IGI Global Scientific Publishing, 2010, pp. 398–423.

[18] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.

[19] K. Wang et al., "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 700–717.

[20] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Trans. Signal Inf. Process.*, vol. 3, no. 1, Nov. 2014, Art. no. e12.

[21] K. Ezzameli and H. Mahersia, "Emotion recognition from unimodal to multimodal analysis: A review," *Inf. Fusion*, vol. 99, no. 1, May 2023, Art. no. 101847.

[22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, no. 1, pp. 98–125, Sep. 2017.

[23] Y. Zhang, Z.-R. Wang, and J. Du, "Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition," in *Proc. Int. Jt. Conf. Neural Netw.*, Budapest, Hungary, 2019, pp. 1–8.

[24] N. Jadhav, "Hierarchical weighted framework for emotional distress detection using personalized affective cues," *J. Inf. Syst. Telecommunication*, vol. 10, no. 38, pp. 89–101, Apr. 2022.

[25] X. Wang, X. Chen, and C. Cao, "Human emotion recognition by optimally fusing facial expression and speech feature," *Signal Process. Image Commun.*, vol. 84, no. 1, May 2020, Art. no. 115831.

[26] S. Zhang et al., "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Syst. Appl.*, vol. 237, no. 1, Sep. 2023, Art. no. 121692.

[27] L. Cai, J. Dong, and W. Min, "Multi-modal emotion recognition from speech and facial expression based on deep learning," in *Proc. Chin. Automat. Congr.*, Shanghai, China, 2020, pp. 5726–5729.

[28] K. Aghajani, "Audio-visual emotion recognition based on a deep convolutional neural network," *J. AI Data Min.*, vol. 10, no. 4, pp. 529–537, Nov. 2022.

[29] Y. L. Bouali, O. B. Ahmed, and S. Mazouzi, "Cross-modal learning for audio-visual emotion recognition in acted speech," in *Proc. Int. Conf. Adv. Technol. Signal Image Process.*, Sfax, Tunisia, 2022, pp. 1–6.

[30] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowl.-Based Syst*, vol. 244, no. 1, May 2022, Art. no. 108580.

[31] A. Al-Saffar, H. Tao, and M. A. Talab, "Review of deep convolution neural network in image classification," in *Proc. Int. Conf. Radar Antenna Microw. Electron. Telecommunication*, Jakarta, Indonesia, 2018, pp. 26–31.

[32] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. K. Wong, and W. Woo, "Convolutional LSTM Network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[33] H. Lian et al., "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, Oct. 2023, Art. no. e25101440.

[34] M. Maithri et al., "Automated emotion recognition: Current trends and future perspectives," *Comput. Methods Programs Biomed.*, vol. 215, no. 1, Jan. 2022, Art. no. 106646.

[35] A.-L. Cîrneanu, D. Popescu, and D. Iordache, "New trends in emotion recognition using image analysis by neural networks, A systematic review," *Sensors*, vol. 23, no. 16, Aug. 2023, Art. no. 7092.

[36] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Inf. Fusion*, vol. 95, no. 1, pp. 306–325, Jul. 2023.

[37] L. Lai, N. Suda, and V. Chandra, "Not all ops are created equal!," in *Proc. Annu. Conf. Mach. Learn. Syst.*, 2018, pp. 987–999.

[38] T.-J. Yang, Y. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6071–6079.

[39] B. Schuller et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, Makuhari, Chiba, Japan, 2010, pp. 2794–2797.

[40] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN," *Sensors*, vol. 17, no. 7, Jul. 2017, Art. no. s17071694.

[41] M. Al-Qaderi, E. Lahamer, and A. Rad, "A two-level speaker identification system via fusion of heterogeneous classifiers and complementary feature cooperation," *Sensors*, vol. 21, no. 15, Jul. 2021, Art. no. s21155097.

[42] F. Wet et al., "Evaluation of formant-like features on an automatic vowel classification task," *J. Acoustical Soc. Amer.*, vol. 116, no. 3, pp. 1781–1792, Sep. 2004.

[43] D. O'Shaughnessy, "Trends and developments in automatic speech recognition research," *Comput. Speech Lang.*, vol. 83, no. 1, Jan. 2024, Art. no. 101538.

[44] N. Singh, R. A. Khan, and R. Shree, "MFCC Adn prosodic feature extraction techniques: A comparative study," *Int. J. Comput. Appl.*, vol. 54, no. 1, pp. 9–13, Sep. 2012.

[45] Z. Liu et al., "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, no. 1, pp. 271–280, Jan. 2018.
[46] Z. Yang, Z. Li, S. Zhou, L. Zhang, and S. Serikawa, "Speech emotion recognition based on multi-feature speed rate and LSTM," *Neurocomputing*, vol. 601, no. 1, Oct. 2024, Art. no. 128177.
[47] S. J. Joysingh, P. Vijayalakshmi, and T. Nagarajan, "Significance of chirp MFCC as a feature in speech and audio application," *Comput. Speech Lang.*, vol. 89, no. 1, Jan. 2025, Art. no. 101713.
[48] P. Ruan, X. Zheng, Y. Qiu, and Z. Hao, "A binaural MFCC-CNN sound quality model of high-speed train," *Appl. Sci.*, vol. 12, no. 23, Nov. 2022, Art. no. 12151.
[49] D. O'Shaughnessy, "Review of analysis methods for speech applications," *Speech Commun*, vol. 151, no. 1, pp. 64–75, Jun. 2023.
[50] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, "Speaker identification features extraction methods: A systematic review," *Expert Syst. Appl.*, vol. 90, no. 1, pp. 250–271, Dec. 2017.
[51] J. Robert, "pydub: Manipulate audio with an simple and easy high level interface," 2018. [Online]. Available: https://pypi.org/project/pydub/
[52] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Kauai, HI, USA, 2003, pp. 511–518.
[53] S. Cheng and G. Zhou, "Facial expression recognition method based on improved VGG convolutional neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 34, no. 7, Oct. 2019, Art. no. 2056003.
[54] T. Debnath et al., "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," *Sci. Rep.*, vol. 12, no. 1, Apr. 2022, Art. no. 6991.
[55] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 11030–11039.
[56] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016.
[57] L. Lu and A. Hanjalic, "Audio segmentation," in *Encyclopedia of Database Systems*, 1st Ed., Boston, MA, USA: Springer, 2009, pp. 167–172.
[58] K. A. Araño, P. Gloor, C. Orsenigo, and C. Vercellis, "When old meets new: Emotion recognition from speech signals," *Cogn. Comput.*, vol. 13, pp. 771–783, Apr. 2021.
[59] X. Shao and S. G. Johnson, "Type-II/III DCT/DST algorithms with reduced number of arithmetic operations," *Signal Process.*, vol. 88, no. 6, pp. 1553–1564, Jun. 2008.
[60] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.
[61] B. McFee et al., "librosa: Audio and music signal analysis in Python," in *Proc. Python Sci. Conf.*, 2015, pp. 18–24.
[62] Z. Kh. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.
[63] M. G. De Pinto, M. Polignano, P. Lops, and G. Semeraro, "Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients," in *Proc. IEEE Conf. Evol. Adaptive Intell. Syst.*, Bari, Italy, 2020, pp. 1–5.
[64] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Composing general audio representation by fusing multilayer features of a pre-trained model," in *Proc. Eur. Signal Process. Conf.*, Belgrade, Serbia, 2022, pp. 200–204.
[65] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv.1609.04747*.
[66] D. P. Kingma and J. A. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representation*. San Diego, CA, USA, May 2015, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1412.6980
[67] J. S. Garofolo et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Philadelphia, PA, USA: Linguistic Data Consortium, 2015, doi: 10.35111/17gk-bn40.
[68] OpenCV, "OpenCV library," 2019. [Online]. Available: https://opencv.org/
[69] U. Michelucci, "Model validation and selection," in *Fundamental Mathematical Concepts For Machine Learning in Science*, Cham, Switzerland: Springer, 2024, pp. 153–184.
[70] M. Wegrzyn, M. Vogt, B. Kireclioglu, J. Schneider, and J. Kissler, "Mapping the emotional face. How individual face parts contribute to successful emotion recognition," *PLoS ONE*, vol. 12, no. 5, May 2017, Art. no. e0177239.
[71] F. W. Smith and S. Rossit, "Identifying and detecting facial expressions of emotion in peripheral vision," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0197160.
[72] A. Diwan, R. Sunil, P. Mer, R. Mahadeva, and S. P. Patole, "Advancements in emotion classification via facial and body gesture analysis : A survey," *Expert Syst.*, vol. 42, Oct. 2024, Art. no. e13759.
[73] M. Kunz, K. Prkachin, P. E. Solomon, and S. Lautenbacher, "Faces of clinical pain: Inter-individual facial activity patterns in shoulder pain patients," *Eur. J. Pain*, vol. 25, no. 3, pp. 529–540, Mar. 2021.
[74] A. K. Anderson, D. D. Spencer, R. K. Fulbright, and E. A. Phelps, "Contribution of the anteromedial temporal lobes to the evaluation of facial emotion," *Neuropsychol*, vol. 14, no. 4, pp. 526–536, 2000.
[75] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Quart. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, Nov. 2010.
[76] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition," in *Proc. IEEE Int. Conf. Automat. Face Gesture Recognit.*, Nara, Japan, 1998, pp. 366–371.
[77] J. Kumari, R. Rajesh, and K. M. Pooja, "Facial expression recognition: A survey," *Procedia Comput. Sci.*, vol. 58, no. 1, pp. 486–491, 2015.
[78] T. Leinonen et al., "Empirical investigation of multi-source cross-validation in clinical ECG classification," *Comput. Biol. Med.*, vol. 183, no. 1, Dec. 2024, Art. no. 109271.
[79] V. Rybalkin, J. Ney, M. K. Tekleyohannes, and N. I., "When massive GPU parallelism ain't enough: A novel hardware architecture of 2D-LSTM neural network," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 15, no. 1, Nov. 2021, Art. no. 2.
[80] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, Kyoto, Japan, 2017, pp. 61–72.
[81] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomed. Signal Process. Control*, vol. 78, Sep. 2022, Art. no. 103970.
[82] M. Sharafi, M. Yazdchi, and J. Rasti, "Audio-visual emotion recognition using K-means clustering and spatio-temporal CNN," in *Proc. Int. Conf. Pattern Recognit. Image Anal.*, Qom, Iran, 2023, pp. 1–6.
[83] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135559–135570, 2021.
[84] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to Human computer interaction," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshop*, Madison, WI, USA, 2003, pp. 53–53.
[85] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Fourth Quarter 2014.
[86] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets," *Eng. Appl. Artif. Intell.*, vol. 130, Apr. 2024, Art. no. 107708.
[87] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach Learn.*, 2017, pp. 3319–3328.
[88] H. Eisenbarth and G. W. Alpers, "Happy mouth and sad eyes: Scanning emotional facial expressions," *Emotion*, vol. 11, no. 4, pp. 860–865, Aug. 2011.
[89] S. D. Gunnery and J. A. Hall, "The expression and perception of the duchenne smile," in *The Social Psychology of Nonverbal Communication*, London. U.K: Palgrave Macmillan, 2015, pp. 114–133.
[90] M. Krumnikl and V. Maiwald, "Facial emotion recognition for mobile devices: A practical review," *IEEE Access*, vol. 12, pp. 15735–15747, 2024.
[91] "Quantization aware training | TensorFlow model optimization," (n.d.). [Online]. Available: https://www.tensorflow.org/model_optimization/guide/quantization/training

Su Yen Ding received the BEng degree (Hons.) in electrical and electronics engineering from Universiti Teknologi PETRONAS, Seri Iskandar, Perak, Malaysia, in 2021. She is currently working toward the MSc degree in electrical and electronics engineering with Universiti Teknologi PETRONAS. Her research interests include emotion recognition and lightweight CNN for real-time implementation.

Tong Boon Tang (Senior Member, IEEE) received the BEng (Hons.) and PhD degrees from the University of Edinburgh. He is currently a professor and the dean of Engineering Faculty with Universiti Teknologi PETRONAS. His research interests include neurotechnology, from device and measurement to data fusion. He received the Lab on Chip Award, in 2006, the IET Nanobiotechnology Premium Award, in 2008, the IET Mountbatten Medal, in 2020, and the Top Research Scientists Malaysia, in 2021. He served as the secretary for the Higher Centre of Excellence (HICoE) Council and the chair for the IEEE Circuits and Systems Society Malaysia Chapter. He is a fellow of IET (U.K.) and a Chartered Engineer.

Cheng-Kai Lu (Senior Member, IEEE) received the PhD degree in engineering from the University of Edinburgh, U.K., in 2012. He is currently a faculty member with the Electrical Engineering Department, National Taiwan Normal University, Taiwan. Previously, he was a faculty member with Universiti Teknologi PETRONAS, Malaysia, from 2016 to 2021. His research focuses on tailored-made artificial intelligence solutions. With more than 8 years of industrial experience, He has published extensively, including book chapters, journal articles, conference papers, and holds several patents.