IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS 2025)
23–24 October 2025, Kushtia, Bangladesh

Attention-Based Deep Learning for Scalable Speech Emotion Recognition with Synthetic Bone-Conducted Speech

Md. Ibne Shihab Shad
Dept. of ICE, PUST
Pabna, Bangladesh
ibn.shihab17@gmail.com

Sayeeda Khan
Dept. of ICE, PUST
Pabna, Bangladesh
sayeedakhan2002@gmail.com

Md. Sarwar Hosain
Dept. of ICE, PUST
Pabna, Bangladesh
sarwar.ice@pust.ac.bd

Akif Mahdi
Dept. of ICE, PUST
Pabna, Bangladesh
akif2100@gmail.com

Manob Chandra Chanda
Dept. of ICE, PUST
Pabna, Bangladesh
manob.210624@s.pust.ac.bd

Md. Rifat Hossain
Dept. of ICE, PUST
Pabna, Bangladesh
rifat.pust.ice14@gmail.com

Abstract—**Speech emotion recognition (SER) is a critical technology that supports advances in human-computer interaction, mental health assessment, and personalized education. Despite significant progress, conventional SER methods relying on air-conducted (AC) speech remain vulnerable to environmental noise and recording inconsistencies, which impede robust real-world deployment. A promising alternative, bone-conducted (BC) speech, captures internal vocal tract vibrations and demonstrates inherent resilience to ambient noise. However, the scarcity of authentic BC datasets and the prohibitive cost of BC sensors severely constrain its practical utilization in SER research. Furthermore, existing efforts have yet to establish scalable approaches to harness BC speech characteristics without specialized acquisition hardware, limiting the generalizability of current models. To address these challenges, we propose a novel framework that synthesizes BC-like speech from standard AC recordings via a carefully designed Infinite Impulse Response (IIR) filter. This method enables cost-effective, large-scale augmentation of training data, circumventing the need for specialized BC sensors. The core of our approach is an attention-augmented convolutional neural network that effectively integrates local spectral feature extraction with long-range temporal modeling. To counteract class imbalance and improve generalization, the model employs class-weighted loss combined with label smoothing. Evaluated on the benchmark RAVDESS dataset, our framework achieves state-of-the-art results: a validation accuracy of 95.51%, balanced accuracy of 95.32%, Matthews Correlation Coefficient of 0.9487, and flawless class-wise AUC scores. A high mean Intersection over Union (IoU) of 0.9163 further confirms the model's precision and robustness across diverse emotional categories.**

Index Terms—**speech emotion recognition, synthetic bone-conducted speech, attention mechanisms, deep learning, noise-robust classification.**

I. Introduction

Emotions are central to human experience, directly influencing behavior, cognition, learning, and mental well-being. In domains such as education and mental health, understanding emotional states plays a critical role—helping personalize learning environments, supporting emotional regulation, and enabling early detection of psychological distress. Among all modalities available for emotion analysis, speech remains the most natural, accessible, and widely interpretable medium [1]. Unlike facial expressions or physiological signals, speech inherently conveys rich emotional cues through variations in pitch, tone, intensity, and rhythm. These prosodic features make it an effective input for machine-based emotion understanding, especially in real-world, non-intrusive settings. Speech emotion recognition (SER) has thus emerged as a key component of affective computing, enabling machines to infer human emotions from acoustic signals [2]. Traditionally, SER systems have relied on air-conducted (AC) speech, which is captured using conventional microphones. However, AC speech is often susceptible to external noise, environmental reverberation, and speaker–microphone variability, which degrade performance and hinder deployment in uncontrolled environments. In contrast, bone-conducted (BC) speech, which captures vibrations transmitted through the skull, exhibits superior robustness to ambient noise and may better reflect internal vocal tract dynamics tied to emotional expression [3]. Despite these advantages, the adoption of real BC speech in SER research is limited by the scarcity of BC sensors, which are often expensive, specialized, and not readily available at scale. To bridge this gap, recent studies have proposed synthetic BC speech generation by transforming AC recordings through signal processing techniques that mimic the spectral behavior of true BC signals—most notably, the suppression of high-frequency components via filtering. This research builds upon that direction by investigating the use of synthetically generated BC speech for emotion recognition. The objective is to mitigate both the noise vulnerabilities of AC speech and the cost / accessibility barriers to true BC acquisition. Inspired by recent findings that demonstrate the effectiveness of BC speech in improving SER performance, this work proposes

979-8-3315-5525-2/25/$31.00 ©2025 IEEE

a hybrid learning framework. The model emphasizes local acoustic features using attention-enhanced convolutional layers and employs class-weighted training with label smoothing to address data imbalance and improve generalization. In doing so, this work contributes a scalable and robust solution for SER, grounded in the benefits of synthetic BC speech modeling that can enhance the reliability of emotion recognition systems in real-world applications.

II. Literature Review

SER has increasingly used deep learning to model complex emotional cues embedded within speech signals. Among publicly available datasets, the *Ryerson Audio-Visual Database of Emotional Speech and Song* (RAVDESS) stands out due to its standardized recording conditions, professional actor performances, and multi-modal structure. It includes utterances expressing eight distinct emotional states: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted, making it a widely used benchmark for SER systems.

Several studies have explored advanced neural architectures on RAVDESS. **Jalal et al. (2019)** reported an impressive **95.1%** using a convolutional neural network (CNN) with self-attention mechanisms [4]. Their input features included F0, MFCCs, energy features with delta coefficients, log-spectrograms (128 filterbanks), eGeMAPS, and supervector representations. However, such high performance raises concerns regarding reproducibility and generalization, especially considering RAVDESS's class imbalance and speaker-dependent nature. In reality, this level of accuracy is difficult to replicate consistently across different experiments and may reflect overfitting to speaker identity rather than robust emotional content. In contrast, **Assunção et al. (2019)** applied speaker embeddings extracted via VGGVox, followed by PCA/LDA for dimensionality reduction, and trained tree-based classifiers, achieving a more conservative maximum accuracy of **81.1%**. This further underscores the challenges posed by data imbalance and speaker variability in surpassing 90% accuracy reliably [5].

A novel direction was introduced by **Hosain et al. (2023)**, who investigated *bone-conducted (BC) speech* for SER [6]. Given the scarcity of real BC speech data, they synthesized BC-like signals from AC speech using an *IIR filter* mimicking bone conduction characteristics. Using *log-Mel spectrograms* and a *CNN with data augmentation*, their model achieved **72.5% accuracy** on synthetic BC speech—surpassing the **69.83%** obtained on original AC speech. This suggests that BC transformation enhances retention of emotionally salient low-frequency components often attenuated in conventional AC recordings. Physiological evidence supports the notion that BC speech better reflects internal vocal tract dynamics, which are closely tied to emotional expression. Moreover, BC-based systems may offer advantages in noisy environments where traditional microphones struggle.

TABLE I: Performance comparison of selected studies on the RAVDESS dataset

Study	Year	Model	Input Features	Acc. (%)
Hosain et al.	2023	CNN	Log-Mel spectrogram (synthetic BC)	**72.5**
Hosain et al.	2023	CNN	Log-Mel spectrogram (original AC)	69.83
Jalal et al.	2019	CNN + SA	F0, MFCC, log-spec., eGeMAPS	**95.1**
Assunção et al.	2019	LMT	VGGVox embeddings + PCA/LDA	81.1

Given the size and imbalance of the RAVDESS data set, consistently achieving very high classification accuracy remains challenging, as confirmed by the notable performance gap between studies like Jalal et al. and Hosain et al. Our work is therefore inspired by the methodology of Hosain et al., aiming to improve SER performance in BC-transformed speech by employing an attention-based CNN architecture optimized for extracting discriminative local features, along with mechanisms to capture global temporal dependencies. We also incorporate class-weighted loss and label smoothing techniques to mitigate class imbalance and enhance model generalization, targeting more robust and reproducible emotion recognition results.

III. Methodology

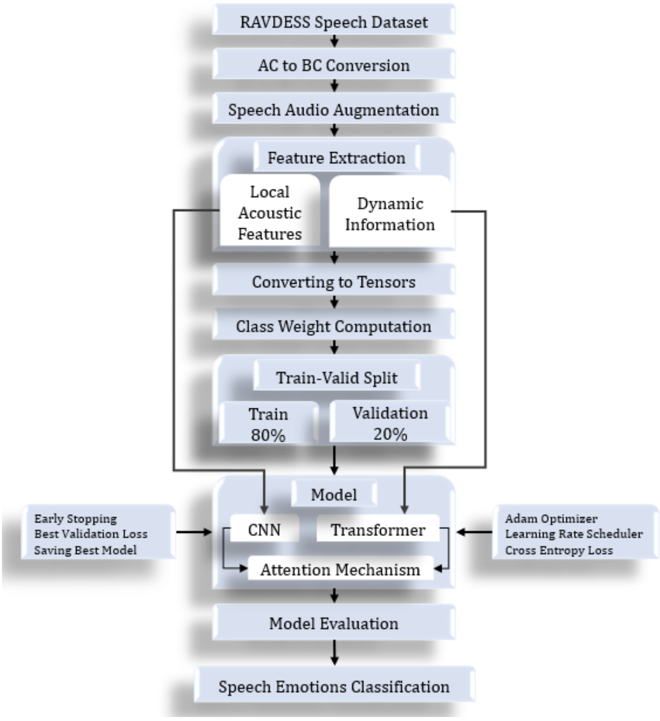

Fig. 1: System workflow block diagram.

A. Dataset description

The RAVDESS is a professionally curated emotional speech corpus designed for affective computing research. For this work, we employed the audio-only speech subset, consisting of 1,440 high-resolution recordings sampled at 48 kHz in 16-bit mono WAV format. The dataset features 24 trained North

American English speakers (balanced by sex), each delivering two neutral sentences in eight distinct emotional categories: neutral, calm, happy, sad, angry, fearful, disgust and surprised.

All expressions were recorded at two emotional intensities (normal and strong), except for the neutral class, which is limited to a single level. The recordings maintain perceptual clarity and acoustic consistency, making them ideal for tasks such as emotion recognition, cross-modal learning, and enhancement of speech signals. The emotional class distribution is balanced, ensuring unbiased training and evaluation of affect-sensitive models [7].

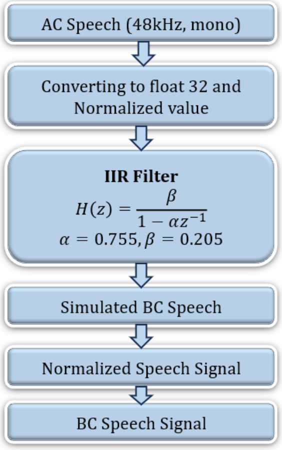
AC Speech (48kHz, mono)
Converting to float 32 and Normalized value
IIR Filter
$H(z) = \frac{\beta}{1 - \alpha z^{-1}}$
$\alpha = 0.755, \beta = 0.205$
Simulated BC Speech
Normalized Speech Signal
BC Speech Signal

Fig. 2: Steps in AC to BC conversion.

B. AC-to-BC speech conversion framework

Synthesizing BC speech from AC speech is essential for applications like silent speech interfaces and hearing assistive devices. While AC speech is widely available, BC speech remains limited due to specialized acquisition requirements.

We propose a simple yet effective linear filtering approach for AC-to-BC conversion. The transformation is modeled using a first-order infinite impulse response (IIR) filter, which captures the low-pass spectral characteristics typical of BC speech—particularly its attenuation of high-frequency components.

Let $A(z)$, $B(z)$, and $H(z)$ denote the z-transforms of the input AC signal, output BC signal, and system transfer function, respectively:

$$B(z) = H(z)A(z) \tag{1}$$

The IIR filter is defined as:

$$H(z) = \frac{\beta}{1 - \alpha z^{-1}} \tag{2}$$

with constraints $0 < \alpha < 1$ and $\beta > 0$ ensuring stability and low-pass behavior. Based on prior analysis, we use averaged filter coefficients $\alpha = 0.755$ and $\beta = 0.205$, enabling robust synthesis without gender-specific tuning.

This streamlined model supports efficient generation of realistic BC-like speech from standard AC data, facilitating data augmentation and enhancing emotion recognition performance under BC conditions [6].

C. Data augmentation for model robustness and future scalability

To simulate real-world acoustic variability and strengthen model generalization, a comprehensive set of audio augmentation techniques was employed across temporal, spectral, amplitude, and spatial domains [6]. These augmentations—including time stretching, pitch shifting, speed perturbation, noise injection, and echo addition—were applied independently with randomized probability, thereby maintaining the semantic integrity of emotional speech while introducing controlled variability into the training distribution.

This augmentation strategy offers two critical advantages. First, it substantially increases the volume of training data, a prerequisite for deploying high-capacity models such as Transformers, which are known to require extensive datasets for optimal convergence [8]. Second, by exposing the model to a wide range of acoustic distortions, it enhances resilience to real-world noise and harsh recording conditions—a vital capability for robust emotion recognition in unconstrained environments.

Through this dual-purpose approach, the augmented dataset not only improves CNN performance but also establishes a scalable framework suitable for future integration with more expressive, data-intensive architectures.

TABLE II: Summary of audio augmentation techniques applied

#	Technique	Category	Details / Parameters
1	Time Stretching	Temporal	Rate = 1.2 (speed up without pitch change)
2	Speed Perturbation	Temporal	Rate = 1.1 (global speed increase)
3	Time Shifting	Temporal	Shift = ±20% of signal length
4	Time Masking	Temporal	Max mask = 10% of audio
5	Pitch Shifting	Spectral	n_steps = +2 semitones
6	Volume Scaling	Amplitude	Gain = 0.5–1.5×
7	Noise Addition	Amplitude	Noise level = 0.005
8	Clipping Distortion	Amplitude	Threshold = ±0.5
9	Polarity Inversion	Amplitude	Inverts waveform polarity
10	Echo Addition	Reverb/Spatial	Delay = 0.5 s, Decay = 0.5
11	Dynamic Compression	Amplitude	Threshold = –20 dB, Ratio = 4:1
12	Vocal Tract Perturb.	Spectral	$\alpha = 1.1$

D. Feature extraction and representation

After augmentation, all speech signals were uniformly resampled to 16 kHz, ensuring consistent temporal resolution while reducing redundant frequency content [9-10]. To capture the nuanced emotional signatures embedded within speech, we extracted a rich set of local acoustic features that reflect short-term spectral energy fluctuations and prosodic dynamics. These features are critical for modeling the transient, emotion-dependent variations in voice tone, pitch, and articulation.

The extracted features were then transformed into structured tensor representations with dimensions $(T \times F)$, where T denotes the temporal frames and F represents the feature dimensionality. This tensor format preserves the sequential and spatial relationships across time and frequency, serving as a robust input to a CNN. The CNN learns hierarchical abstractions from these tensors, enabling it to detect salient emotional cues across local temporal regions. This end-to-end formulation bridges low-level signal variations with high-level affective representations, forming the backbone of our emotion recognition pipeline.

E. Class imbalance mitigation

While extensive augmentation was performed to expand the dataset, the resulting emotional class distribution remained inherently imbalanced—most notably, the neutral category comprises only half the number of samples relative to other emotional classes. Such skewed representation can adversely affect learning dynamics, leading to biased predictions toward majority classes. To address this, we employed a class-weighted loss strategy, wherein inverse frequency-based weights were computed from the training labels using a balanced heuristic. These weights were cast into tensor form and integrated directly into the loss function, ensuring proportional error contribution from all classes during optimization. This adjustment enhances the model's sensitivity to minority classes without artificially oversampling them. The final class distribution motivating this correction is illustrated in Fig. 3.

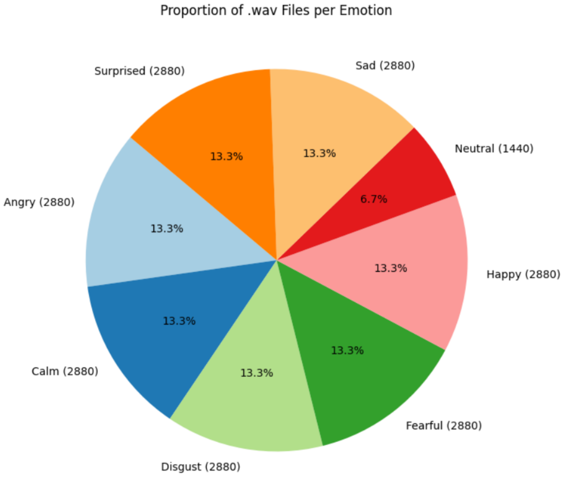

Fig. 3: Emotions distribution in RAVDESS augmented speech dataset.

F. Proposed model and training procedure

The proposed model integrates convolutional layers [11] with a Transformer encoder to effectively extract both local and long-range temporal features from bone-conducted speech for emotion recognition. Convolutional layers capture fine-grained spectral details, while the Transformer models temporal dependencies to highlight emotionally salient segments. An attention mechanism further refines these representations before classification, enabling nuanced differentiation across eight emotion categories.

A key challenge addressed during training is the class imbalance in the augmented dataset, where the Neutral class is significantly underrepresented. To counteract this, class weights were computed using inverse class frequency and incorporated into the cross-entropy loss function as tensor weights. This strategy ensures that minority classes contribute more heavily to the loss, thereby reducing bias toward majority classes and enhancing the model's sensitivity to less frequent emotions.

In addition, label smoothing with a factor of 0.1 was applied to regularize the model's output distribution. By softening hard target labels, label smoothing mitigates overconfidence in predictions, improves generalization, and reduces overfitting risks—critical for modeling the subtle acoustic nuances of emotional speech.

The training dataset was partitioned with an 80:20 split for training and validation. Using a batch size of 64, the model was optimized with the Adam optimizer, leveraging its adaptive learning rates for efficient convergence. Data loading employed a custom PyTorch DataLoader with shuffling to ensure randomized input order each epoch. Training was controlled via early stopping with a patience of 5 epochs monitoring validation loss, leading to convergence typically within 15 to 20 epochs—significantly faster than traditional schedules. This demonstrates the effectiveness of the combined loss strategies and optimizer in stabilizing and accelerating learning.

Overall, this training paradigm, grounded in balanced loss optimization and regularization, enabled the model to robustly learn discriminative emotional features from bone-conducted speech despite class imbalance and acoustic complexity.

IV. Performance Evaluation and Results

Model performance was assessed using key evaluation metrics, including balanced accuracy, Matthews correlation coefficient (MCC), mean squared error (MSE), and log loss, complemented by training–validation trend analysis to ensure convergence stability.

A. Classification report

The model achieved a high overall accuracy of 96%, with consistently strong precision, recall, and F1-scores across all emotion classes. The performance is well-balanced, as reflected in the close alignment of macro and weighted averages, indicating robust generalization and minimal class imbalance effects.

TABLE III: Classification report on the validation set

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Angry	1.00	0.93	0.97	568
Calm	0.98	0.97	0.98	508
Disgust	0.98	0.96	0.97	514
Fearful	0.90	0.99	0.94	607
Happy	0.88	0.97	0.92	615
Neutral	0.98	0.92	0.95	303
Sad	0.97	0.91	0.94	606
Surprised	0.98	0.96	0.97	599
Accuracy		0.96		4320
Macro Avg	0.96	0.95	0.96	4320
Weighted Avg	0.96	0.96	0.96	4320

B. Confusion matrix

The normalized confusion matrix presented below offers a detailed visualization of the model's predictive distribution across emotion classes, highlighting both the accuracy and misclassification patterns inherent in the classification process.

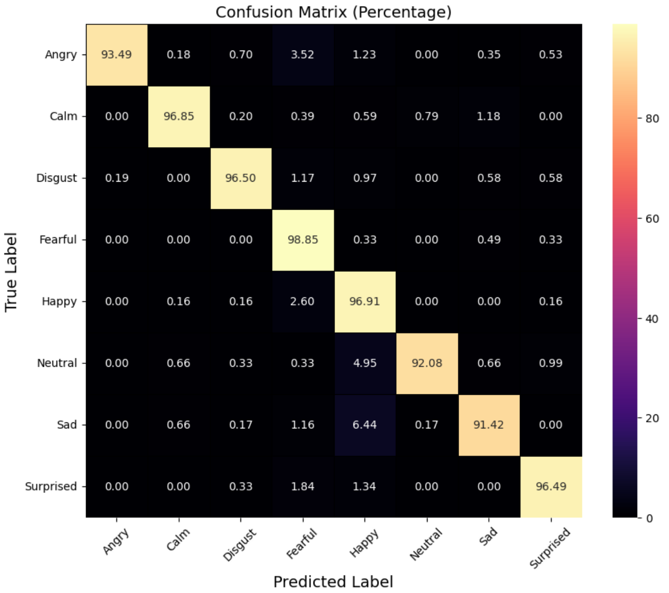

Fig. 4: Confusion matrix of the proposed model on RAVDESS.

C. ROC-AUC curve analysis

The ROC-AUC curve analysis further substantiates the model's discriminative strength, with each class achieving a perfect AUC score of 1.0. This indicates flawless separability between classes, as visually confirmed by the sharply defined ROC curves.

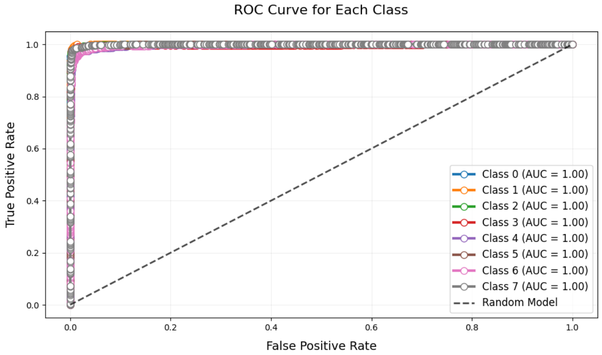

Fig. 5: The ROC-AUC curve of the proposed model on RAVDESS.

D. Intersection over union (IoU) analysis

Although intersection over union (IoU) is conventionally used in segmentation tasks to measure spatial overlap, we repurpose it here to assess class-wise prediction quality in a multi-class classification setting. Computed from the confusion matrix, the IoU for each class c is defined as:

$$\mathrm{IoU}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c + \mathrm{FN}_c} \tag{3}$$

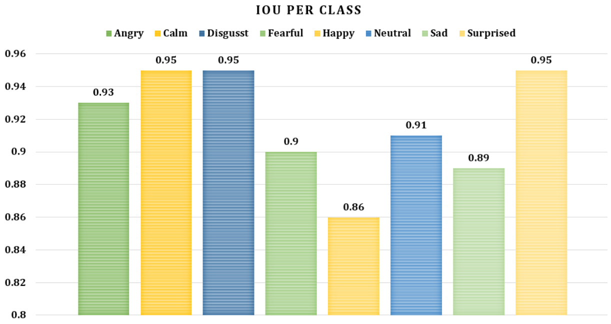

Fig. 6: The classwise IoU scores of the proposed model on RAVDESS.

where TP_c, FP_c, and FN_c represent the true positives, false positives, and false negatives for class c, respectively.

The model achieved a strong *mean IoU* of **0.9163**, reflecting balanced classification performance across all emotion categories. In particular, the *Calm* and *Surprised* classes yielded the highest IoU scores of **0.95**, indicating consistent and highly accurate predictions.

E. Mean squared error analysis

The model yielded a low overall MSE of 0.0812, indicating minimal prediction variance. Among all classes, neutral and disgust emotions exhibited the lowest MSE values of 0.0514 and 0.0769, respectively, suggesting high prediction reliability for these emotional states. The detailed class-wise error distribution is illustrated above.

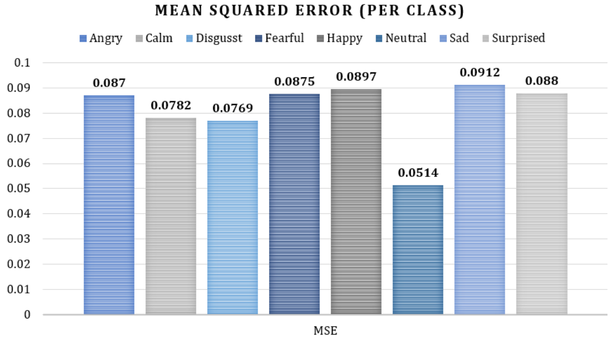

Fig. 7: The classwise MSE of the proposed model on RAVDESS.

F. Logarithmic loss analysis

The model achieved an overall logarithmic loss of 1.4105, reflecting a strong confidence alignment between predicted and true class probabilities. Class-wise log loss values remain consistently low, with the Neutral emotion showing the best performance at 0.2074, indicating highly calibrated predictions. The complete distribution is visualized in Figure8.

G. Training and validation performance

The training–validation accuracy trends indicate that the model generalized effectively, with no significant divergence

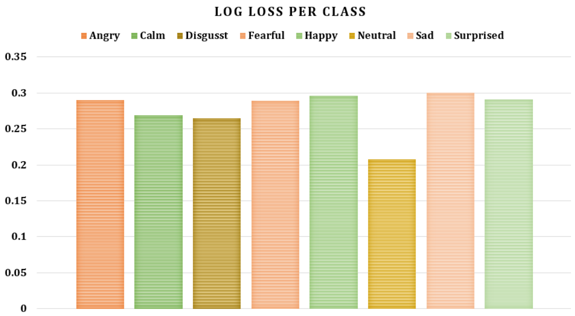

Fig. 8: The classwise logarithmic loss of the proposed model on RAVDESS.

observed between the training and validation accuracy curves throughout the learning process. This close alignment suggests stable learning and minimal overfitting. Moreover, an early stopping criterion with a patience of 5 was applied, leading the training to halt automatically at the 22nd epoch—well before any signs of performance degradation or overfitting emerged. This not only preserved model generalizability but also optimized computational efficiency. The corresponding training dynamics are illustrated in the figure below.

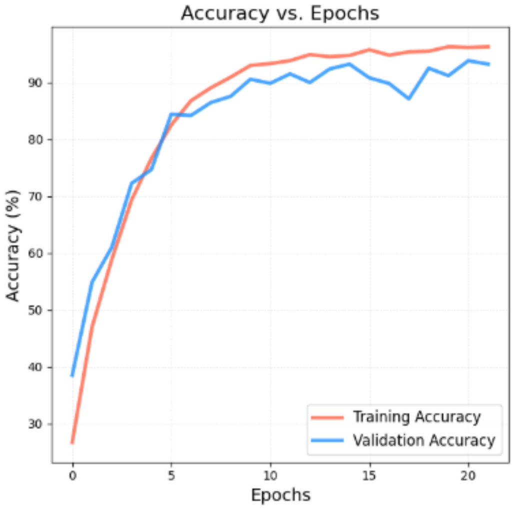

Fig. 9: The Training-Validation accuracy graph of the proposed model.

H. Proposed model performance evaluation

The proposed model demonstrates strong and reliable performance, as reflected by its final training and validation accuracies of 94.50% and 95.51%, respectively, with low corresponding loss values, indicating stable convergence and no overfitting. The high MCC of 94.87% further confirms the model's robust predictive power, particularly in handling multi-class imbalances. Additionally, a balanced accuracy of 95.32% highlights its consistency in correctly identifying samples across all emotion categories, regardless of class distribution. Collectively, these metrics underscore the proposed model's effectiveness, generalizability, and practical applicability in real-world emotion recognition tasks.

TABLE IV: Proposed model performance evaluation

Metric	**Value**
Final Training Accuracy	94.50%
Final Validation Accuracy	95.51%
Balanced Accuracy	95.32%
Matthews Correlation Coefficient (MCC)	0.9487
Final Training Loss	0.6226
Final Validation Loss	0.6019

V. CONCLUSION

The proposed emotion recognition model demonstrates strong generalization, high accuracy, and consistent performance across all evaluation metrics. With an overall accuracy of 95.51%, a high MCC of 94.87%, and a balanced accuracy of 95.32%, the model proves robust even in the presence of inter-class variability. Additional analyses—such as per-class AUC (1.0 for all), high Mean IoU (0.9163), and low MSE and Log Loss—further validate the model's reliability and confidence calibration. The use of early stopping ensured efficient convergence without overfitting, as confirmed by closely aligned training and validation curves. Collectively, these results affirm the effectiveness and practical applicability of the proposed model in real-world emotion recognition tasks.

REFERENCES

[1] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, “Detection of clinical depression in adolescents' speech during family interactions,” IEEE Transactions on Biomedical Engineering, vol. 58, no. 3, pp. 574–586, 2010.

[2] S. Kwon: A CNN-assisted enhanced audio signal processing for speech emotion recognition, Sensors, Vol. 20, p. 1838, 2020.

[3] Y. Zhou, Y. Chen, Y. Ma Y and H. Liu: A real-time dualmicrophone speech enhancement algorithm assisted by bone conduction sensor, MDPI Sensors, Vol. 20, No. 18, p. 5050, 2020.

[4] Jalal, M. A., Moore, R. K., & Hain, T. (2019). Spatio-temporal context modelling for speech emotion classification. In 2019 IEEE automatic speech recognition and understanding workshop (ASRU) (pp. 853–859). IEEE.

[5] Assunção, G., Menezes, P., & Perdigão, F. (2019). Importance of speaker specific speech features for emotion recognition. In 2019 5th experiment international conference (exp. at'19) (pp. 266–267). IEEE

[6] M. S. Hosain, Y. Sugiura, N. Yasui, and T. Shimamura, "Deep-Learning-Based Speech Emotion Recognition Using Synthetic Bone-Conducted Speech," *Journal of Signal Processing*, vol. 27, no. 6, pp. 151–163, Nov. 2023.

[7] 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, pp. 1–35, May 2018. doi: 10.1371/journal.pone.0196391

[8] L. Tarantino, P. N. Garner, A. Lazaridis et al., “Self-attention for speech emotion recognition,” in Interspeech, 2019, pp. 2578–2582

[9] Deng, J., Zhang, Z., Eyben, F., & Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. IEEE Signal Processing Letters, 21, 1068–1072.

[10] Shih, P.-Y., Chen, C.-P., & Wang, H.-M. (2017). Speech emotion recognition with skewrobust neural networks. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2751–2755). IEEE.

[11] Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA) (pp. 1–4).