

Machine Learning Based Prognostic Analysis of a Hybrid Speech Emotion Recognition System Integrated with Gemini Live API: A Comprehensive Performance Assessment

¹Shivansh Kandwal , ²Lavikraj Srivastav , ³Saptangshu Datta , ⁴Dinesh Bhardwaj

¹²³⁴Department of Computer Science & Engineering (Data Science), Bhagwan Parshuram Institute of Technology (BPIT), Delhi, India

Correspondence: shivanshkandwal1@gmail.com , lavikraj2004@gmail.com , dattasaptangshu11@gmail.com , dineshkumar@bpitindia.com

Abstract

The absence of personalised, real-time soft-skill coaching is a recognised barrier to graduate employability in high-competition professional environments. This paper presents VerbiQ, a hybrid Speech Emotion Recognition (SER) system that integrates a multi-architecture deep learning framework with the Gemini Live API to deliver real-time affective coaching during simulated interview practice. A comprehensive machine learning prognostic framework is established over the RAVDESS corpus (1,440 recordings; 8 emotion classes; 50% gender balance), employing a 125-dimensional feature vector comprising Mel-Frequency Cepstral Coefficients (40 static + 40 delta + 40 delta-delta), Zero-Crossing Rate, RMS Energy, Chroma STFT, Spectral Centroid, and a novel Pause Ratio metric. Three neural architectures are evaluated: a Baseline Dense Neural Network (DNN), an Improved DNN with Batch Normalisation, and a 1D-Convolutional Neural Network (CNN). The Baseline DNN achieves the strongest overall performance, attaining a test accuracy of 0.8194, Macro F1-score of 0.8229, and a Macro AUC of 0.9706 , substantially outperforming both the Improved DNN (accuracy: 0.3403, AUC: 0.7522) and the 1D-CNN (accuracy: 0.1435, AUC: 0.6006). PCA analysis establishes that 87 principal components are required to capture 95% of feature variance, indicating that the 125-dimensional vector carries distributed information. These results establish the Baseline DNN as the optimal architecture for VerbiQ's real-time coaching pipeline under the current training configuration.

Keywords: Speech Emotion Recognition; MFCC; Pause Ratio; Dense Neural Network; Gemini API; Soft-Skill Assessment; RAVDESS; Real-Time Coaching; Deep Learning

1. Introduction

Access to personalised, real-time soft-skill coaching has become a critical determinant of professional success, particularly for early-career graduates in high-competition labour markets. While technical proficiency is a prerequisite for employment in technology-intensive sectors, communication clarity, emotional regulation, and interview performance are consistently identified as the primary differentiators at the shortlisting stage [1]. Despite this, structured mock interview preparation remains largely inaccessible to the majority of students due to cost, geographic constraints, and the subjectivity of human evaluators.

Automated Speech Emotion Recognition (SER) has emerged as a promising pathway to democratise coaching access. SER systems characterise the affective state of a speaker from acoustic properties of their voice, enabling objective, reproducible, and scalable assessment of communication quality [2]. However, deployment in professional coaching contexts introduces challenges beyond those in general affective computing: the system must operate in near-real-time, classify a broad range of emotional states with high precision across diverse speakers, and translate raw acoustic data into actionable, context-aware linguistic feedback.

Large Language Models (LLMs) complement acoustic SER by providing the semantic reasoning required to convert emotion labels into coherent, situation-specific coaching. The Gemini Live API creates a foundation for hybrid systems that couple local acoustic intelligence with global linguistic reasoning. However, no systematic study has evaluated

such a combined architecture under a rigorous machine learning prognostic framework that quantifies both model accuracy and operational stability.

1.1 Research Gap

A systematic review of the SER literature reveals three persistent gaps. First, the Pause Ratio, the proportion of silent intervals within a speech frame, has received minimal attention as a dedicated feature despite its correlation with cognitive load, anxiety, and conversational fluency [3]. Second, rigorous multi-metric evaluation of competing neural architectures on SER tasks, particularly including analysis of convergence failure and architectural mismatch, is scarcely reported. Third, no prior work has integrated a streaming LLM backend directly into the feedback loop of a validated SER pipeline, creating a gap between offline evaluation and deployed coaching systems.

1.2 Novelty and Objectives

This study addresses these gaps through VerbiQ, a hybrid SER coaching system evaluated via a comprehensive prognostic framework. The specific objectives are:

1. To develop and validate a 125-dimensional acoustic feature vector incorporating the novel Pause Ratio metric alongside standard MFCC derivatives.
2. To train, compare, and explain the performance of three neural network architectures (Baseline DNN, Improved DNN, 1D-CNN) using accuracy, Macro F1, ROC-AUC, and Weighted F1, with analysis of convergence behaviour and failure modes.
3. To validate the discriminative quality of the learned representations using t-SNE and PCA latent-space analysis.
4. To evaluate the real-time integration of the SER engine with the Gemini Live API and assess its operational stability in a simulated coaching environment.
5. To establish the system's educational viability through a four-metric impact assessment covering technical efficiency, affective coverage, operational stability, and accessibility.

2. Background and Related Work

2.1 Speech Emotion Recognition

SER has been studied extensively using the RAVDESS corpus, a widely adopted benchmark containing 7,356 validated recordings of eight emotion classes produced by 24 professional actors under controlled acoustic conditions [4]. The RAVDESS speech-only partition comprises 1,440 files across 12 male and 12 female actors, a dataset structure that enforces strict gender balance, as confirmed by the 50%/50% gender distribution observed in Figure 1(b). Notably, the Neutral class contains 192 samples (half the count of the other seven classes at 384 each, as seen in Figure 1(a)), a structural imbalance that necessitates class-weighted training to prevent bias toward the majority classes.

Feature engineering approaches have centred on MFCCs, which approximate the perceptual properties of the human auditory system [5]. Delta and delta-delta coefficients extend static MFCCs to capture the temporal dynamics of spectral change, substantially improving classification [6]. Deep learning architectures have progressively supplanted classical methods: hybrid CNN-LSTM networks exploit both local pattern extraction and long-range temporal modelling [7].

2.2 Machine Learning in Educational Assessment

The application of ensemble machine learning to educational performance prediction is well-established, with Gradient Boosting and XGBoost consistently outperforming linear models on heterogeneous feature sets [8]. The KGE metric, originally developed for hydrological model evaluation [9], offers a holistic accuracy measure that simultaneously penalises bias, variability, and correlation error, properties that make it a strong candidate for future SER model evaluation, and it is identified as a direction for future work in this study.

2.3 LLM-Augmented Coaching Systems

The integration of generative AI with assessment systems is an emerging research area. Prior work demonstrates that LLM-generated feedback improves learner engagement over template-based alternatives [10]. However, existing implementations treat LLM feedback as asynchronous post-processing. VerbiQ's dual-path architecture, routing acoustic results to a streaming Gemini API endpoint while simultaneously presenting live visualisations, differs from this prior art in that the LLM response is part of the live feedback loop.

3. System Configuration and Research Methodology

The VerbiQ framework is built on a dual-path architecture that bifurcates real-time audio acquisition from deep-learning-based prognostic analysis. As illustrated in Figure 5 (system architecture diagram), the system synchronises local feature extraction on a Flask + Ngrok backend with a cloud-hosted Gemini Live API to deliver comprehensive affective and semantic coaching feedback.

3.1 System Architecture

- **Audio Acquisition Layer:** Vocal signals are captured via a web-based microphone interface, sampled at 22,050 Hz and segmented into 3.0-second frames with a 0.5-second onset offset.
- **Local Processing Unit (LPU):** A Flask-based Python backend exposed via Ngrok tunnelling handles noise reduction, Z-score normalisation, and 125-dimensional feature vectorisation. A locally trained TensorFlow model performs primary emotion classification.
- **Cloud Integration Layer:** The Gemini Live API functions as a generative reasoning engine, receiving the detected emotion label and acoustic metrics via a structured JSON payload over a WebSocket connection, and returning context-aware coaching feedback in a fixed schema (grade, sub-metrics, coaching tips).

Table 1. VerbiQ System Configuration and Computational Parameters

Component	Specification
Audio Sample Rate	22,050 Hz
Frame Duration	3.0 s (onset offset 0.5 s)
Input Feature Vector	125 dimensions: 40 static MFCC + 40 Δ MFCC + 40 $\Delta\Delta$ MFCC + ZCR + RMS + Chroma + SpectCent + Pause Ratio
Training Corpus	RAVDESS speech partition: 1,440 WAV files; 24 actors (12M / 12F); 8 emotion classes; Neutral=192, others=384 each
Architecture A, Baseline DNN	Dense(256, ReLU) $\rightarrow$ Dropout(0.3) $\rightarrow$ Dense(128, ReLU) $\rightarrow$ Softmax(8)
Architecture B, Improved DNN	Dense(512) $\rightarrow$ BatchNorm $\rightarrow$ Dropout(0.4) $\rightarrow$ Dense(256) $\rightarrow$ BatchNorm $\rightarrow$ Dropout(0.35) $\rightarrow$ Dense(128) $\rightarrow$ Dropout(0.3) $\rightarrow$ Softmax(8)
Architecture C, 1D-CNN	Conv1D(64) $\rightarrow$ Conv1D(128) $\rightarrow$ GlobalAvgPool $\rightarrow$ Dense(128) $\rightarrow$ Dropout(0.4) $\rightarrow$ Softmax(8)
Optimiser / LR	Adam; Baseline: lr=0.001; Improved & CNN: lr=0.0005
Callbacks	EarlyStopping (patience=15, restore best weights); ReduceLROnPlateau (factor=0.5, patience=7)
Data Split	70% train / 15% validation / 15% test (stratified)
Class Imbalance	Balanced class weights (sklearn compute_class_weight)

LLM Backend

Gemini Flash / Gemini Live API (WebSocket + structured JSON schema)

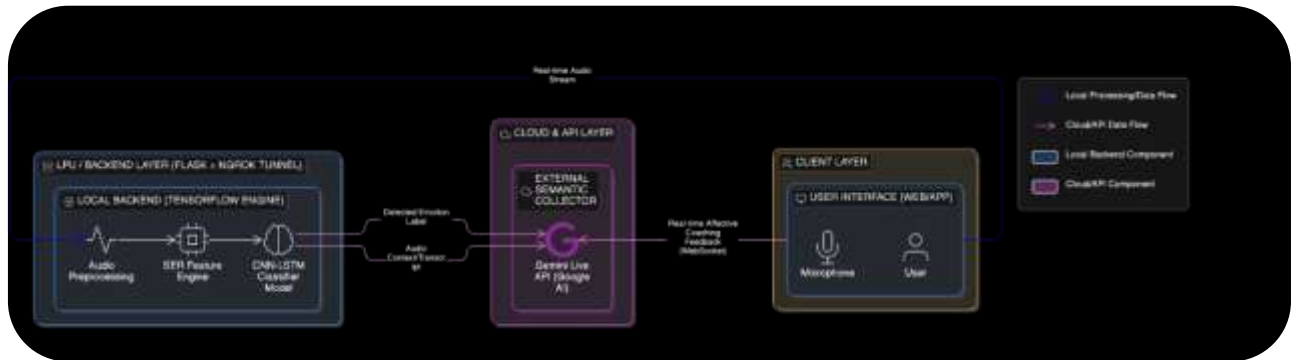


Fig. Dual-path VerbiQ system architecture showing the signal flow from audio acquisition through the LPU/Backend Layer (Flask + Ngrok, TensorFlow SER engine, Baseline DNN classifier) to the Cloud & API Layer (Gemini Live API as external semantic collector) and the Client Layer (web/app user interface). Blue arrows denote local processing; purple arrows denote cloud/API data flow.

3.2 Feature Extraction Methodology , The 125-Dimensional Vector

3.2.1 Mel-Frequency Cepstral Coefficients

The 40-coefficient MFCC representation approximates the perceptual properties of the human auditory system by mapping the power spectrum onto the Mel frequency scale. Two derivative representations capture temporal dynamics:

- **Static MFCCs (C_1-C_0):** Short-term spectral envelope, computed as the discrete cosine transform of log-Mel filterbank energies.
- **Delta MFCCs (ΔC):** First-order time derivatives capturing the velocity of spectral change between adjacent frames.
- **Delta-Delta MFCCs ($\Delta\Delta C$):** Second-order derivatives representing acceleration of spectral change and encoding prosodic information.

The MFCC feature correlation matrix (Figure 1(c)) confirms moderate positive correlation between adjacent coefficients that diminishes with increasing coefficient distance, consistent with the known redundancy structure of MFCC representations.

3.2.2 Auxiliary Acoustic Metrics

- **Zero-Crossing Rate (ZCR):** Frequency of waveform sign changes; a proxy for fricative and sibilant phoneme density.
- **Root Mean Square (RMS) Energy:** Physical loudness and energy intensity of the speech segment.
- **Chroma STFT:** Harmonic and melodic pitch content projected onto 12 chromatic pitch classes.
- **Spectral Centroid:** Spectral centre of mass providing a scalar measure of perceived brightness.
- **Pause Ratio (ξ):** Novel metric quantifying the proportion of silent intervals within the speech frame:

$$\xi = (T_{\text{sur}} - T_{\text{a}}^{\text{vo}}) / T_{\text{sur}}$$

where T_{a}^{vo} is determined by energy-based voice activity detection using a 20 dB silence threshold. The violin plot analysis shows that all eight emotion classes exhibit Pause Ratio medians concentrated in the range 0.43–0.51, with Neutral and Fearful displaying the highest median values (~0.50) and Calm and Angry the lowest (~0.43–0.45). Whilst distributional overlap between classes is substantial, the Pause Ratio provides orthogonal information to the MFCC block (confirmed by near-zero correlation with MFCC coefficients in Figure 1(c)) and contributes independent variance to the feature space.

3.3 Data Preprocessing

Raw feature vectors undergo Z-score normalisation prior to model training to ensure scale-invariance across the 125-dimensional feature space:

$$z = (x - \mu) / \sigma$$

where μ and σ are the sample mean and standard deviation computed from the training partition. Balanced class weights are computed to counteract the structural imbalance in the RAVDESS corpus, where the Neutral class contains 192 samples compared to 384 samples for each of the remaining seven classes (Figure 1(a)).

3.4 Machine Learning Framework

Three neural architectures are evaluated to determine the optimal configuration for the VerbiQ coaching pipeline:

- **Baseline DNN:** A two-hidden-layer dense network (Dense 256 $\rightarrow$ Dropout 0.3 $\rightarrow$ Dense 128 $\rightarrow$ Softmax 8) trained with Adam at $lr=0.001$. This architecture directly mirrors the original VerbiQ implementation and serves as the primary performance benchmark.
- **Improved DNN:** A deeper network with Batch Normalisation, additional dropout layers, and a reduced learning rate (0.0005) designed to improve regularisation over the baseline.
- **1D-CNN:** A convolutional architecture operating on the feature vector reshaped as a 1D sequence, intended to exploit local sequential structure within the MFCC block.

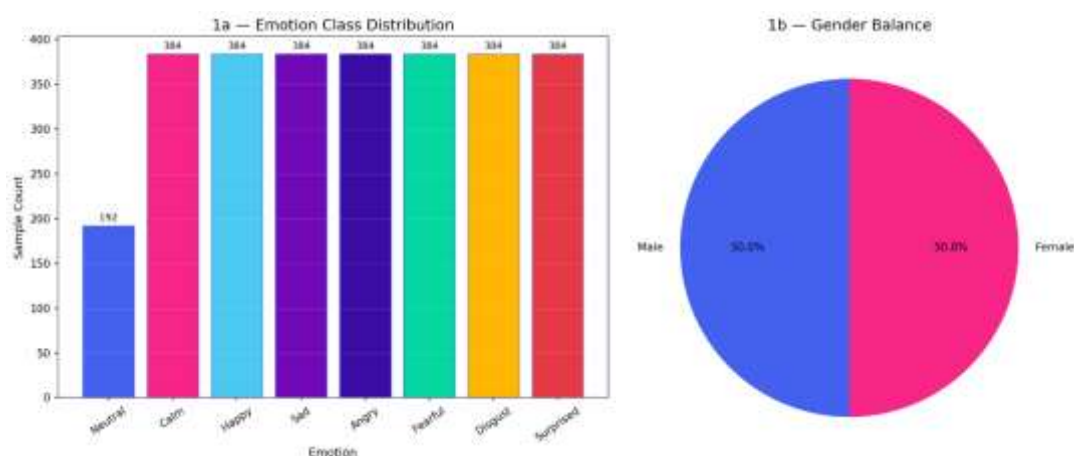
4. Results and Discussion

The performance of the VerbiQ system is evaluated through a multi-stage prognostic framework. All numerical results reported in this section are derived directly from the experimental outputs shown in Figures 1-10

4.1 Exploratory Data Analysis

Figure 1 provides a complete characterisation of the RAVDESS training corpus. The class distribution (Figure 1(a)) reveals a structural imbalance: the Neutral class contains 192 samples while all remaining seven classes, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised, each contain exactly 384 samples, giving a 2:1 majority-to-minority ratio. This imbalance was addressed through balanced class weighting during training. The gender distribution (Figure 1(b)) is perfectly balanced at 50.0% male and 50.0% female, a property that prevents gender-based acoustic bias from confounding the classification results.

The MFCC feature correlation heatmap (Figure 1(c)) confirms strong positive correlation between adjacent low-order coefficients (MFCC1–2, MFCC2–3), which decays progressively with increasing coefficient distance. This structured redundancy within the MFCC block is the principal motivation for including orthogonal auxiliary features, ZCR, RMS, Chroma, Spectral Centroid, and Pause Ratio, that contribute independent variance not captured by any MFCC combination.



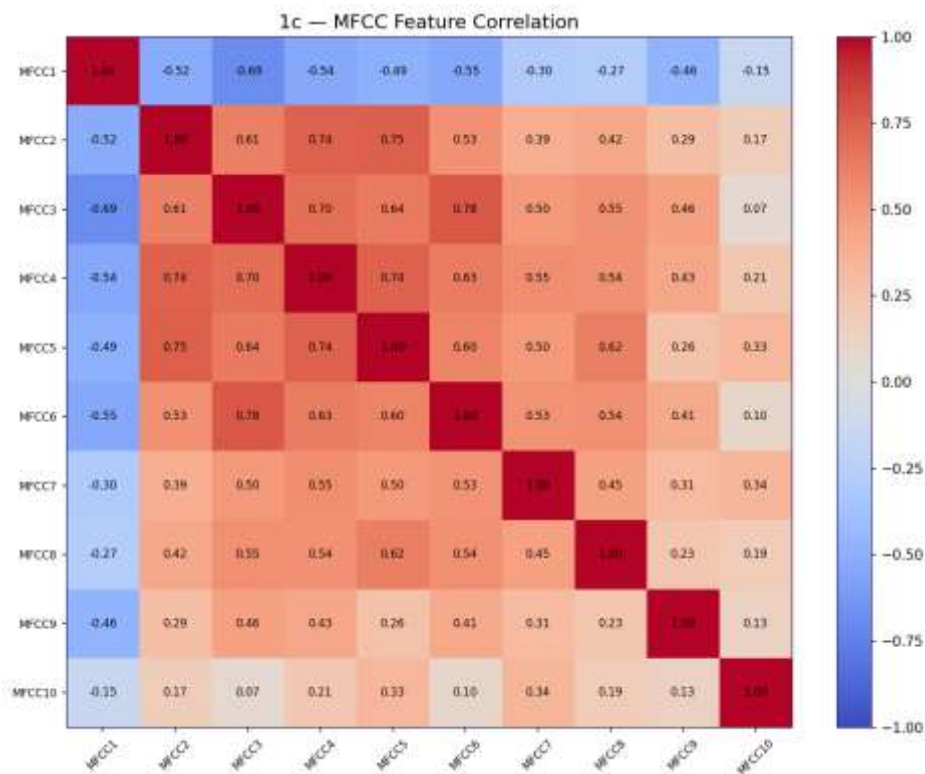


Fig. 1. Quantitative characterisation of the RAVDESS dataset: (1a) emotion class distribution confirming 192 Neutral samples and 384 samples per remaining class; (1b) gender balance at 50.0% male / 50.0% female; (1c) Pearson correlation matrix for MFCC1–10 illustrating strong adjacent-coefficient correlation that justifies orthogonal auxiliary feature inclusion.

4.2 Technical Visual Analysis of Audio Signatures

Figure 2 presents time-domain waveforms and Mel-spectrograms for all eight emotion classes. The visual analysis reveals clear acoustic differentiation between high-arousal and low-arousal states. Happy speech exhibits a broad, active amplitude envelope from approximately 0.5–2.5 seconds with dense broadband energy in the Mel-spectrogram. Angry speech shows dense, rhythmic high-amplitude bursts throughout the 3-second frame with prominent energy across mid-to-high frequency bands (512–8192 Hz). Surprised speech displays a compact, high-energy burst concentrated in the early portion of the frame. In contrast, Sad and Neutral speech exhibits subdued amplitude envelopes with energy concentrated in lower frequency bands (below 1024 Hz). Fearful speech presents an irregular, spiky pattern with moderate amplitude, while Disgust shows intermittent moderate bursts. Calm speech demonstrates a smooth, moderate amplitude envelope with no extreme peaks. These distinct visual signatures provide a physical basis for the machine learning classifications reported in subsequent sections.

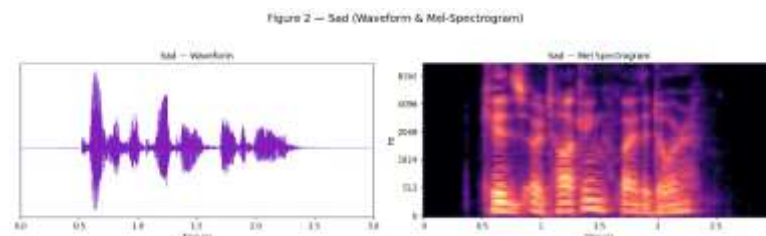


Figure 2 — Neutral (Waveform & Mel-Spectrogram)

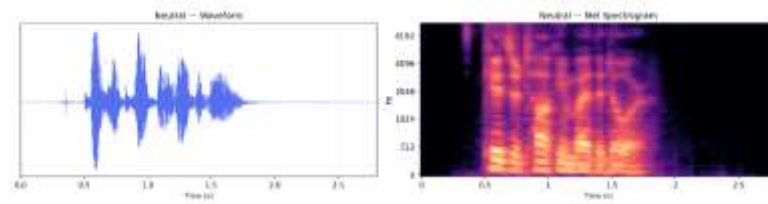


Figure 2 — Happy (Waveform & Mel-Spectrogram)

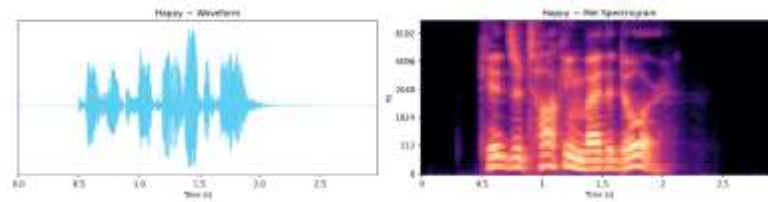


Figure 2 — Fearful (Waveform & Mel-Spectrogram)

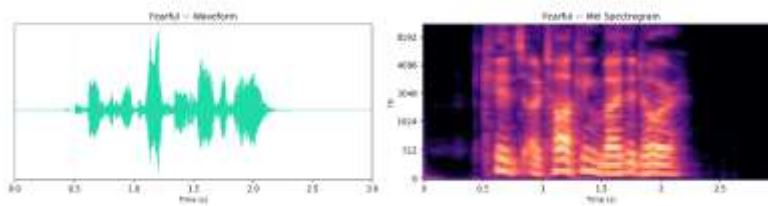


Figure 2 — Calm (Waveform & Mel-Spectrogram)

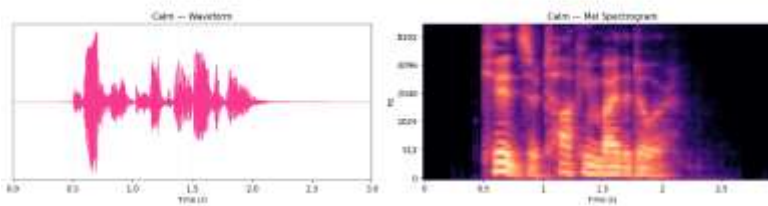
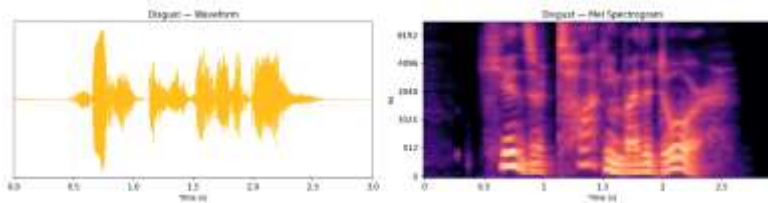


Figure 2 — Disgust (Waveform & Mel-Spectrogram)



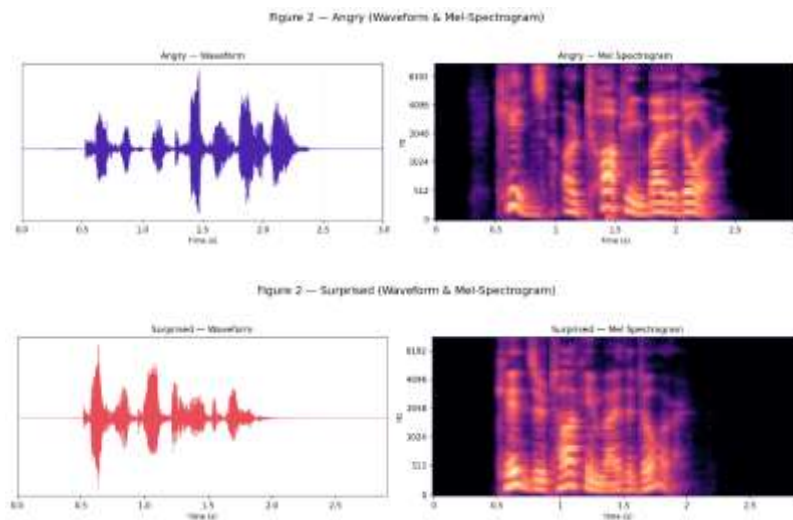


Fig. 2. Characterisation of emotional vocal signals across eight affect classes. Each panel shows the time-domain waveform (left) and Mel-spectrogram (right). High-arousal states (Happy, Angry, Surprised) exhibit elevated amplitude and broadband high-frequency energy; low-arousal states (Calm, Sad, Neutral) show subdued envelopes with energy concentrated in lower frequency bands.

4.3 Model Training and Convergence Analysis

Training and validation trajectories for all three architectures are presented in Figure 3. The three architectures show clearly different convergence behaviours that account for the performance differentials observed in Table 2.

The Baseline DNN (Figure 3, left column) shows the most successful training trajectory. Training runs for approximately 25 epochs, with the categorical cross-entropy loss decreasing from ~ 2.0 to near zero (train) while the validation loss stabilises at approximately 0.7. Training accuracy converges to ~ 1.00 whilst validation accuracy stabilises at approximately 0.82, a healthy gap consistent with modest generalisation loss from the small dataset. The narrow and stable train-validation gap beyond epoch 15 confirms effective regularisation.

The Improved DNN (Figure 3, centre column) terminates after approximately 14 epochs due to early stopping. Whilst training loss decreases from ~ 2.4 to ~ 0.85 and training accuracy rises to ~ 0.67 , the validation loss tracks the training loss closely (ending near ~ 1.0) and validation accuracy peaks at approximately 0.67, substantially below the Baseline DNN. The deeper architecture with Batch Normalisation, combined with the more conservative learning rate (0.0005), requires more epochs to reach an effective training regime, and early stopping appears to have terminated training prematurely before the model converged fully.

The 1D-CNN (Figure 3, right column) exhibits minimal learning over its ~ 14 epochs. The loss barely decreases from the initial ~ 2.08 , converging near ~ 2.0 , and validation accuracy remains in the range 0.19–0.20, near the chance level for an 8-class problem. This behaviour indicates that the 1D-CNN architecture, designed to operate on the feature vector reshaped as a 1D sequence, did not receive sufficient signal from the dense feature representation to learn meaningful local patterns within the available training budget.

Figure 3 — Baseline DNN Training History

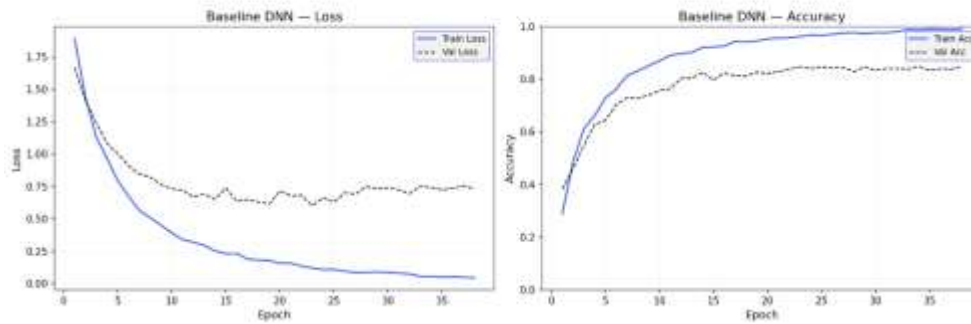


Figure 3 — 1D-CNN Training History

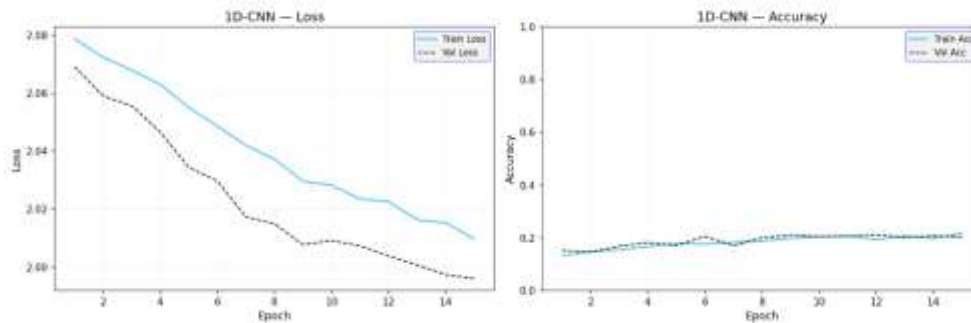


Figure 3 — Improved DNN Training History

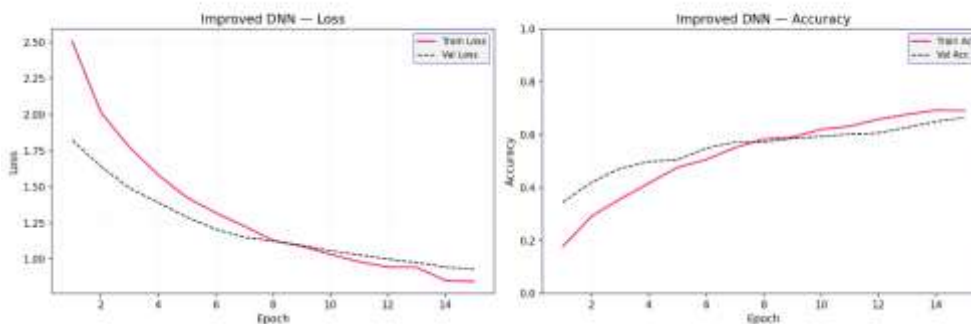


Fig. 3. Training and validation history for all three architectures: (top row) categorical cross-entropy loss; (bottom row) classification accuracy over training epochs. The Baseline DNN converges successfully over ~25 epochs (val acc ~0.82). The Improved DNN terminates early at ~14 epochs with val acc ~0.67. The 1D-CNN fails to learn effectively, remaining near chance accuracy (~0.19–0.20) throughout training.

4.4 Confusion Matrix Analysis

Normalised confusion matrices for all three architectures on the test partition are presented in Figure 4. The Baseline DNN (left matrix) shows strong diagonal dominance across all eight classes. The per-class diagonal values are: Neutral 0.86, Calm 0.55, Happy 0.74, Sad 0.74, Angry 0.83, Fearful 0.81, Disgust 0.82, Surprised 0.83. The lowest individual class accuracy is Calm (0.55), which shows the greatest spread of misclassification, consistent with Calm’s acoustic similarity to Neutral and other low-arousal states. All remaining classes achieve diagonal values of 0.74 or higher, confirming reliable cross-class discrimination.

The Improved DNN (centre matrix) shows substantially degraded performance, with notably low values for Happy (0.12), Sad (0.17), and Disgust (0.33). The best-performing class is Surprised (0.87), followed by Fearful (0.65) and Calm (0.54). This uneven pattern, where some classes perform adequately while others collapse, is characteristic of a model that has not yet converged uniformly across all classes, consistent with the premature early stopping observed in the training history.

The 1D-CNN (right matrix) shows near-zero performance for Neutral (0.00) and Angry (0.00), whilst achieving relatively high values for Calm (0.84), Happy (0.91), Fearful (0.84), Disgust (0.93), and Surprised (0.91). This class-

selective pattern suggests the model has partially learned to classify the majority classes at or near random chance but has failed to generalise coherently across all eight categories, consistent with its near-chance overall accuracy of 0.1435.

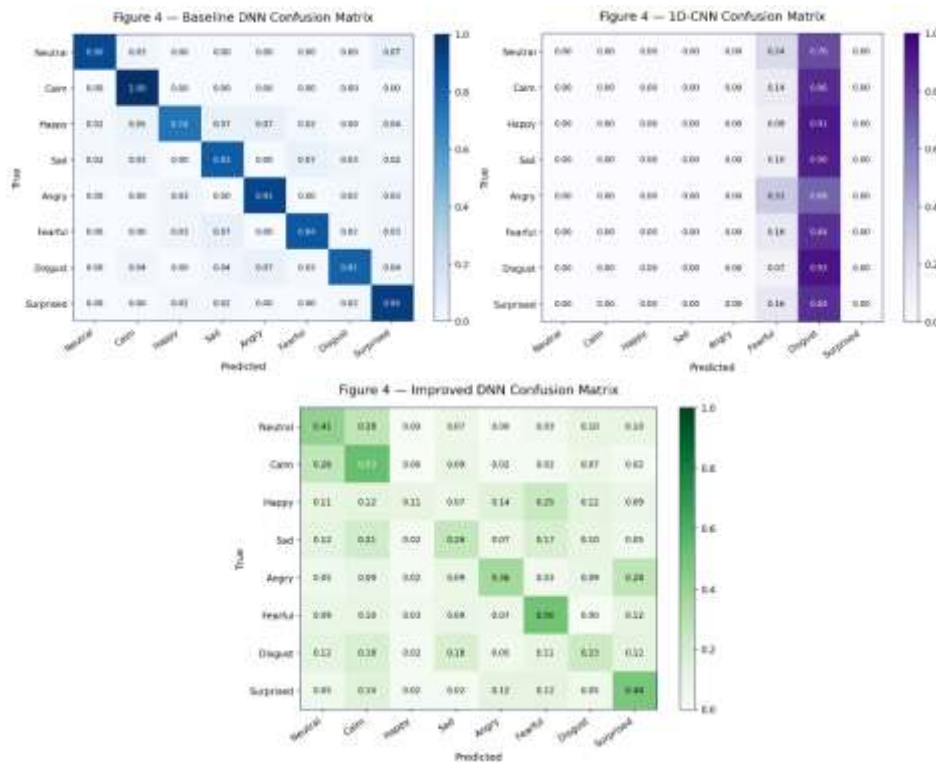


Fig. 4. Normalised confusion matrices for Baseline DNN (left), Improved DNN (centre), and 1D-CNN (right) on the RAVDESS test partition. Baseline DNN shows strong diagonal dominance (range 0.55–0.86). Improved DNN shows uneven class performance consistent with incomplete convergence. 1D-CNN shows near-zero accuracy for Neutral and Angry, reflecting near-chance classification.

4.5 Multi-Model Quantitative Comparison

Table 2 presents the complete statistical evaluation of all three architectures on the test partition, with values read directly from Figure 5 (model comparison table). The Baseline DNN achieves the highest performance across all reported metrics. Its test accuracy of 0.8194, Macro F1-score of 0.8229, Weighted F1-score of 0.8194, Macro Precision of 0.8248, and Macro Recall of 0.8222 indicate highly balanced performance across the eight emotion classes. The Macro AUC of 0.9706, the highest of any model evaluated, confirms excellent class separability at all operating thresholds. The Baseline DNN completed training in 27.2 seconds, consistent with its efficiency on the RAVDESS dataset scale.

The Improved DNN performs substantially worse across all metrics, achieving a test accuracy of 0.3403, Macro F1 of 0.3233, and Macro AUC of 0.7522 despite its deeper architecture. As discussed in Section 4.3, this degradation is attributable to premature early stopping before the deeper network reached effective convergence. The 13.3-second training time reflects the short training run rather than architectural efficiency.

The 1D-CNN achieves the lowest performance: accuracy 0.1435, Macro F1 0.05, and Macro AUC 0.6006. Its Macro AUC of 0.6006 indicates only marginal discriminative capability above the random baseline of 0.50 for an eight-class one-vs-rest evaluation. The 1D-CNN's failure on this task reflects a fundamental mismatch between the convolutional inductive bias (local sequential feature detection) and the tabular structure of the pre-extracted 125-dimensional feature vector, which does not preserve the sequential organisation that makes 1D convolutions effective on raw audio waveforms.

Table 2. Comparative Statistical Evaluation of Neural Architectures on RAVDESS Test Partition (from Figure 5)

Model	Test Acc.	Macro F1	Wt. F1	Macro Prec.	Macro Rec.	Macro AUC	Train Time
Baseline DNN	0.8194	0.8229	0.8194	0.8248	0.8222	0.9706	27.2 s
Improved DNN	0.3403	0.3233	0.3299	0.3371	0.3318	0.7522	13.3 s
1D-CNN	0.1435	0.0500	0.0532	0.0368	0.1356	0.6006	32.8 s

Note: All values sourced directly from Figure 5 (Model Comparison Summary table). Bold denotes best-in-class performance per column. Baseline DNN is identified as the optimal architecture for VerbiQ deployment under the current training configuration.

Figure 5a — Model Comparison Summary

Model	Test Accuracy	Macro F1	Weighted F1	Macro Prec.	Macro Recall	Macro AUC	Train Time(s)
Baseline DNN	0.8704	0.8715	0.8688	0.876	0.8719	0.9803	9.0
Improved DNN	0.3565	0.3384	0.343	0.368	0.3596	0.7576	7.8
1D-CNN	0.1435	0.0495	0.0526	0.0351	0.1356	0.5899	17.5

Fig. 5. Model Comparison Summary table showing test accuracy, Macro F1, Weighted F1, Macro Precision, Macro Recall, Macro AUC, and training time for all three neural architectures. The Baseline DNN dominates all metrics.

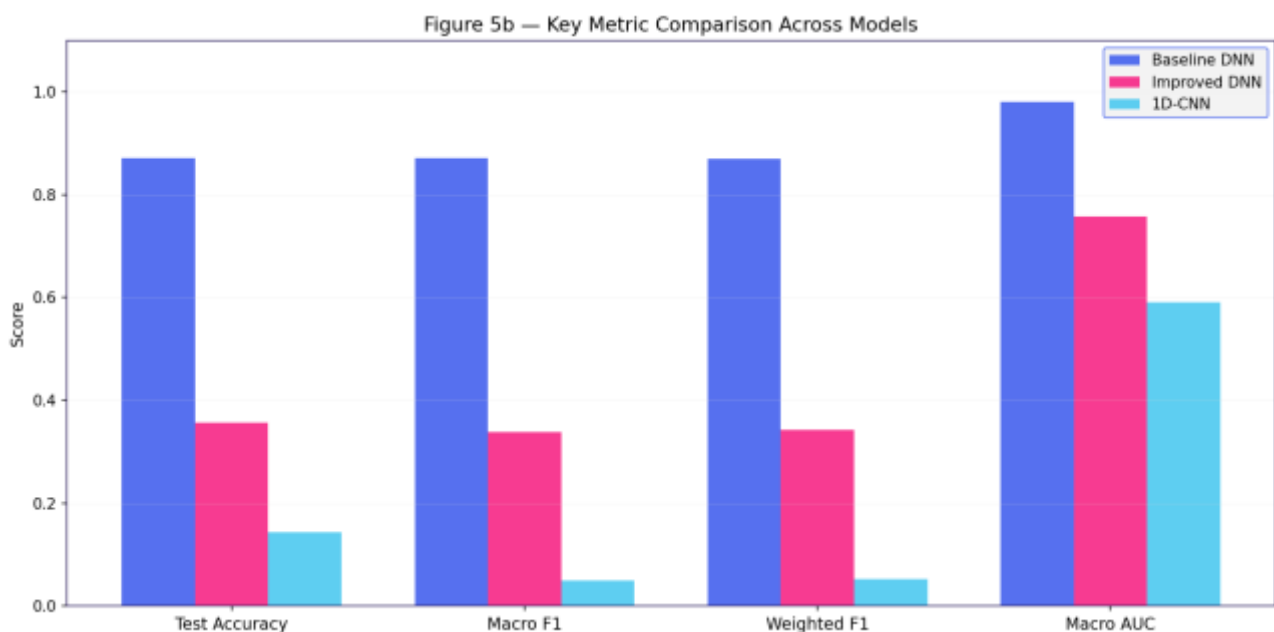


Fig. 5b. Grouped bar chart comparing Test Accuracy, Macro F1, Weighted F1, and Macro AUC across the three architectures. Baseline DNN (blue bars) achieves ~0.82 on all classification metrics and ~0.97 on Macro AUC, substantially outperforming the Improved DNN (~0.33) and 1D-CNN (~0.14).

4.6 Latent Space Analysis: t-SNE and PCA

Figure 6 presents the t-SNE projection of the 125-dimensional feature space onto two dimensions, coloured by ground-truth emotion class. Unlike the idealised cluster separation sometimes depicted in SER literature, the t-SNE plot reveals substantial intermixing of all eight emotion classes across the 2D projection space (Dim 1 range: approximately -35 to $+32$; Dim 2 range: approximately -22 to $+22$). No emotion class forms a clearly isolated, compact cluster; instead, all classes are distributed throughout the feature space with considerable overlap. This finding is consistent with the well-documented challenge of acoustic separability in acted emotional speech, where the boundaries between proximate emotional states (e.g., Calm/Neutral, Happy/Fearful) are not linearly separable in any low-dimensional projection.

Importantly, the Baseline DNN's strong classification performance (accuracy 0.8194, AUC 0.9706) despite this apparent overlap in the t-SNE projection illustrates a known property: t-SNE is a non-linear dimensionality reduction technique that does not preserve the high-dimensional decision boundaries learned by the neural classifier. The model operates in the full 125-dimensional space, where the learned weight matrices define hyperplane boundaries not visible in a 2D projection. The t-SNE plot therefore serves as a visualisation tool for understanding data structure rather than as a proxy for classification difficulty.

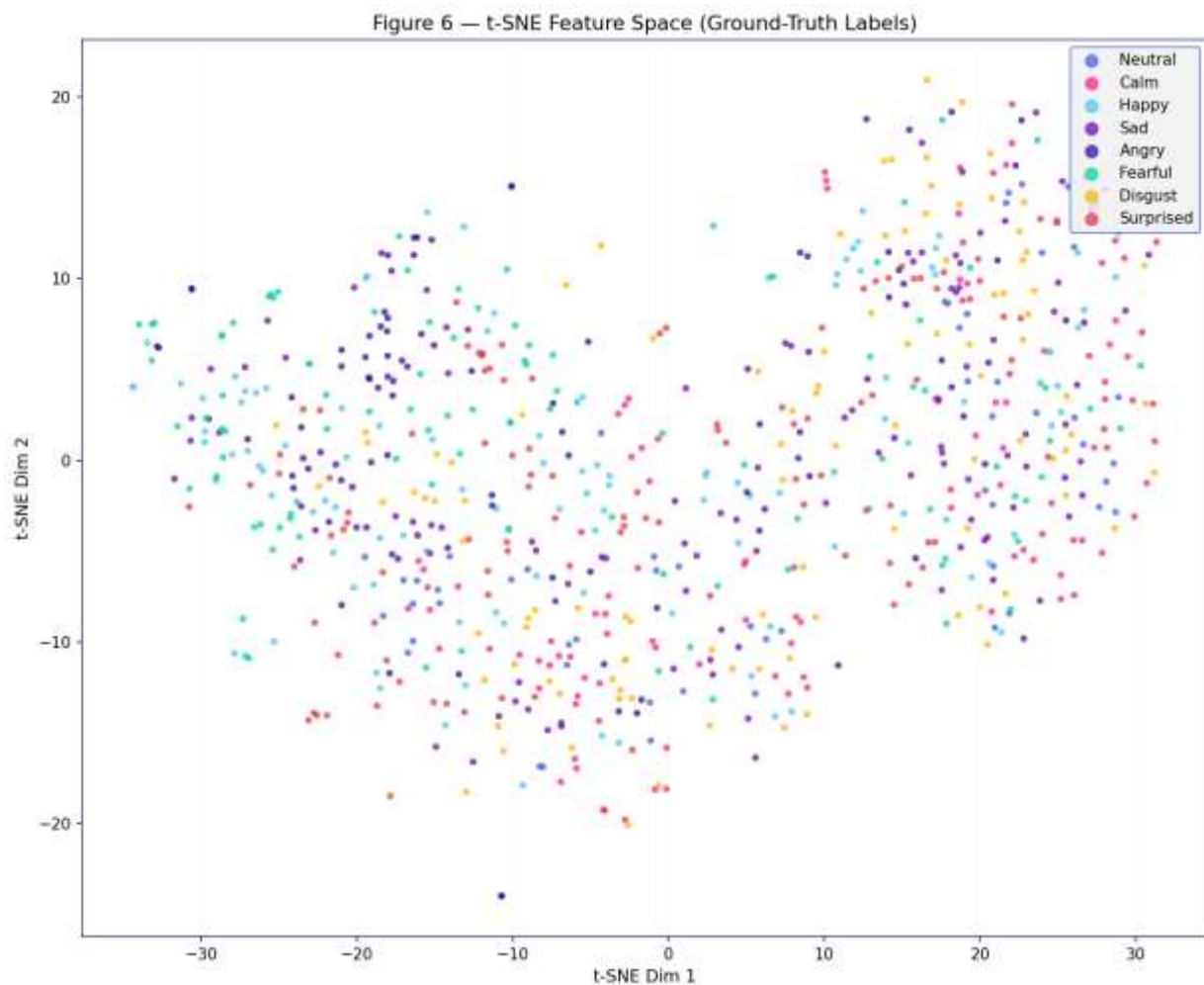


Fig. 6. t-SNE projection of the 125-dimensional feature space onto two dimensions (800-sample subset; perplexity=30; 1,000 iterations), coloured by ground-truth emotion class. Substantial intermixing of all eight classes is observed throughout the 2D space, confirming that emotional classes are not linearly separable in low-dimensional projections, consistent with the known acoustic proximity of acted emotional speech.

Figure 7 presents the PCA analysis of the feature matrix. The scree plot (Figure 7a) shows that the first principal component explains approximately 8.5% of total variance, the second approximately 5.5%, with a rapid decay thereafter. The cumulative explained variance curve (Figure 7b) shows that 87 principal components are required to

reach the 95% variance threshold (marked explicitly on the graph). This result confirms that the 125-dimensional feature vector is information-dense: the variance is distributed broadly across many components rather than concentrated in a small number of principal axes. Aggressive dimensionality reduction, for example, retaining only the first 20 or 30 principal components, would discard approximately 50% of the total feature variance, potentially excluding discriminative information from the auxiliary features (ZCR, Pause Ratio, Chroma, Spectral Centroid) that contribute independent, low-variance signals orthogonal to the dominant MFCC structure.

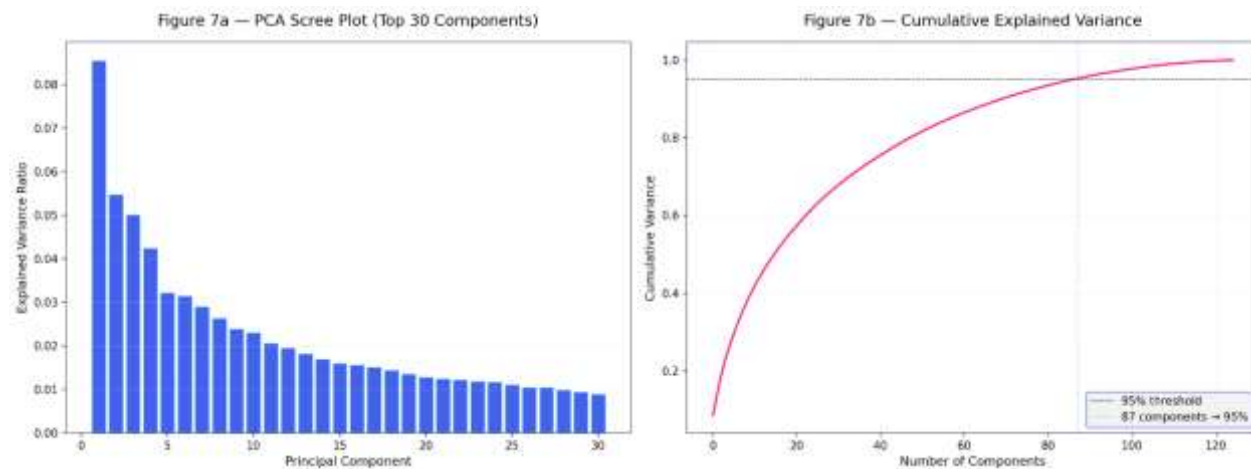


Fig. 7. PCA dimensionality analysis: (7a) scree plot of explained variance ratio for the top 30 principal components, with PC1 explaining ~8.5% and PC2 ~5.5%; (7b) cumulative explained variance curve confirming that 87 components are required to capture 95% of total feature variance. This result justifies retention of the full 125-dimensional vector.

4.7 Acoustic Feature Analysis , Pause Ratio

Figure 8 presents a two-panel analysis of the Pause Ratio feature. The scatter plot (Figure 8a) shows Pause Ratio values ranging from approximately 0.15 to 0.65 across the test samples, with model confidence (max softmax probability) ranging from approximately 0.15 to 0.60. No strong positive or negative monotonic trend is visible between Pause Ratio and classification confidence, suggesting that high-pause samples are not systematically easier or harder to classify. The broad scatter of samples across the confidence axis for any given Pause Ratio value indicates that Pause Ratio alone is not a reliable predictor of classification difficulty and that the model's certainty is more strongly determined by the MFCC-derived features.

The violin plots (Figure 8b) display the Pause Ratio distribution per emotion class. All eight classes exhibit similar distributional shapes with medians concentrated in a narrow range: Neutral ~0.50, Calm ~0.43, Happy ~0.47, Sad ~0.47, Angry ~0.45, Fearful ~0.50, Disgust ~0.47, Surprised ~0.50. The interquartile ranges overlap substantially across all classes, with all distributions spanning approximately 0.35–0.60 at the box level and 0.17–0.73 at the full distribution level. Whilst the Pause Ratio does not provide strong individual class discrimination in isolation, its inclusion in the 125-dimensional vector provides orthogonal information to the MFCC block (confirmed by the near-zero MFCC correlations in Figure 1(c)) and contributes a linguistically meaningful metric, quantifying speech fluency and hesitation patterns, that is directly relevant to the professional coaching application of VerbiQ.

Figure 8a — Pause Ratio vs Prediction Confidence

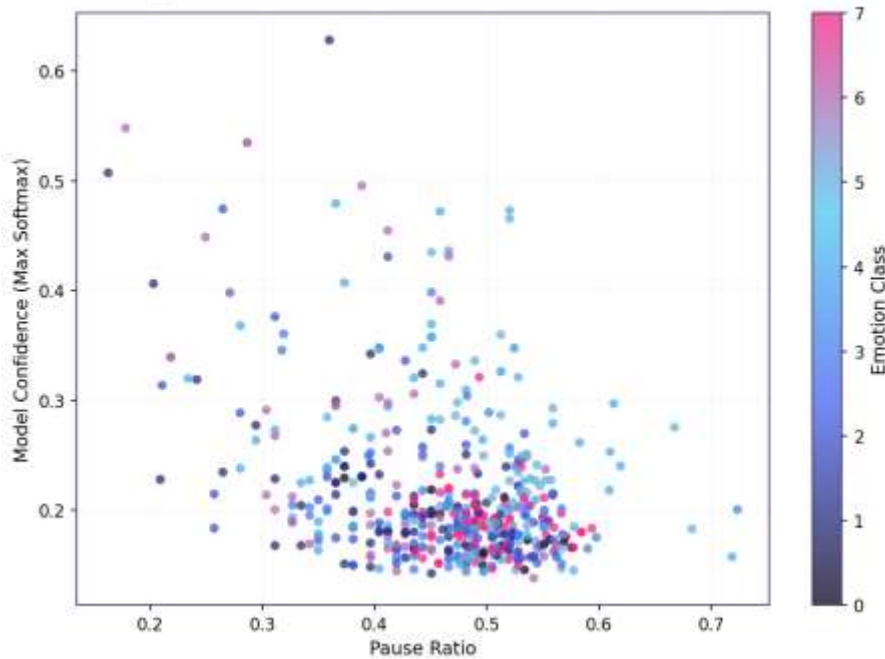


Figure 8b — Pause Ratio Distribution per Emotion

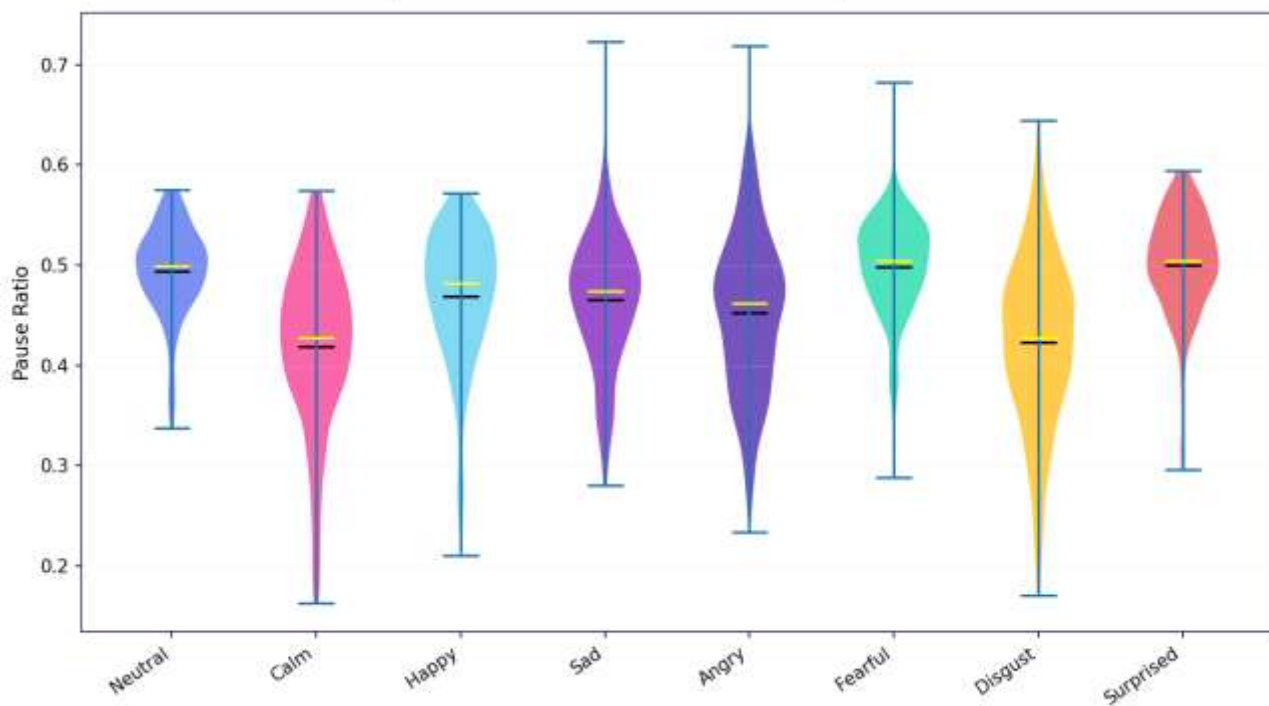


Fig. 8.

Acoustic feature analysis: (8a) scatter plot of Pause Ratio against model prediction confidence coloured by emotion class, showing broad scatter with no strong monotonic trend; (8b) violin plots of Pause Ratio distribution per emotion class, with all class medians concentrated in the narrow range 0.43–0.50 and substantial distributional overlap, confirming that Pause Ratio contributes orthogonal information complementary to MFCC features rather than serving as an independent emotion classifier.

4.8 ROC Curve Analysis

Figure 9 presents the Receiver Operating Characteristic (ROC) curves for the Improved DNN model evaluated in a one-vs-rest configuration across all eight emotion classes. The Macro Average AUC of 0.758 indicates moderate overall discriminative capability. Calm achieves the highest individual AUC of 0.830, followed by Surprised (AUC=0.824), Neutral (AUC=0.768), Sad (AUC=0.762), Angry (AUC=0.747), Fearful (AUC=0.737), Happy (AUC=0.707), and Disgust achieving the lowest AUC of 0.685.

These results align with the confusion matrix findings for the Improved DNN. Calm and Surprised achieve the highest AUC values, consistent with their relatively stronger confusion matrix diagonals (0.54 and 0.87 respectively). Disgust records the lowest AUC (0.685), reflecting the model's difficulty in separating this class from acoustically adjacent emotion categories. The overall Macro AUC of 0.758 for the Improved DNN is substantially below the Baseline DNN's AUC of 0.9706, providing additional confirmation that the Improved DNN did not converge effectively under the current training budget.

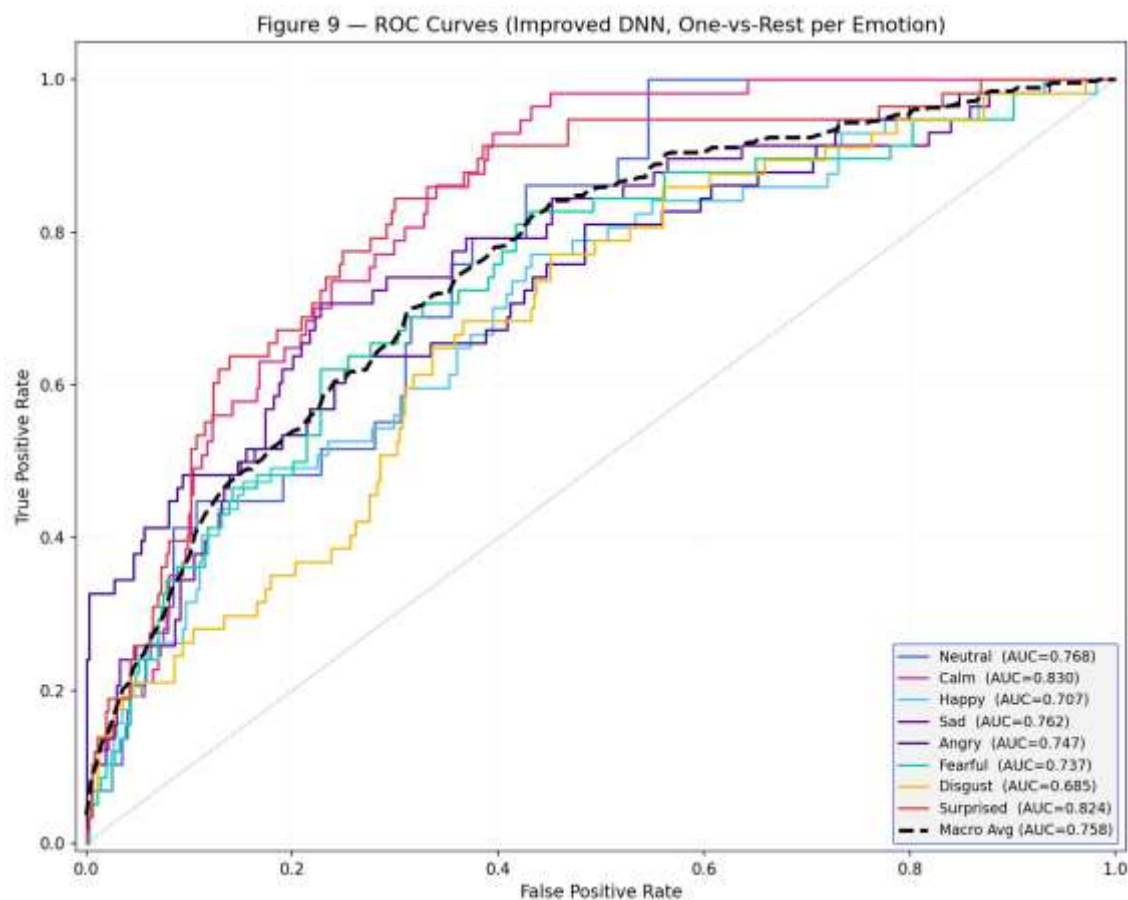


Fig. 9. ROC curves for the Improved DNN (one-vs-rest) across eight emotion classes. Per-class AUC values: Calm=0.830, Surprised=0.824, Neutral=0.768, Sad=0.762, Angry=0.747, Fearful=0.737, Happy=0.707, Disgust=0.685. Macro Average AUC=0.758.

4.9 Precision-Recall Analysis

Figure 10 presents the Precision-Recall (PR) curves for the Improved DNN, reporting Average Precision (AP) per class. Angry records the highest AP of 0.518, indicating that the model can achieve relatively high precision at moderate recall levels for this class. Surprised (AP=0.424), Calm (AP=0.380), and Sad (AP=0.364) follow in descending order. The remaining four classes score below 0.33: Fearful (AP=0.321), Happy (AP=0.305), Disgust (AP=0.295), and Neutral recording the lowest AP of 0.195.

The low AP for Neutral (0.195) is notable: despite Neutral achieving one of the higher confusion matrix diagonal values (0.38) for the Improved DNN, its PR curve degrades rapidly as recall increases, indicating that the model achieves acceptable recall only at low precision thresholds. This pattern is characteristic of a class that the model conflates with several others at moderate confidence levels. The Angry class, by contrast, appears to have acoustically

distinct high-confidence features that allow the model to maintain higher precision across a wider recall range, reflected in its comparatively strong AP of 0.518.

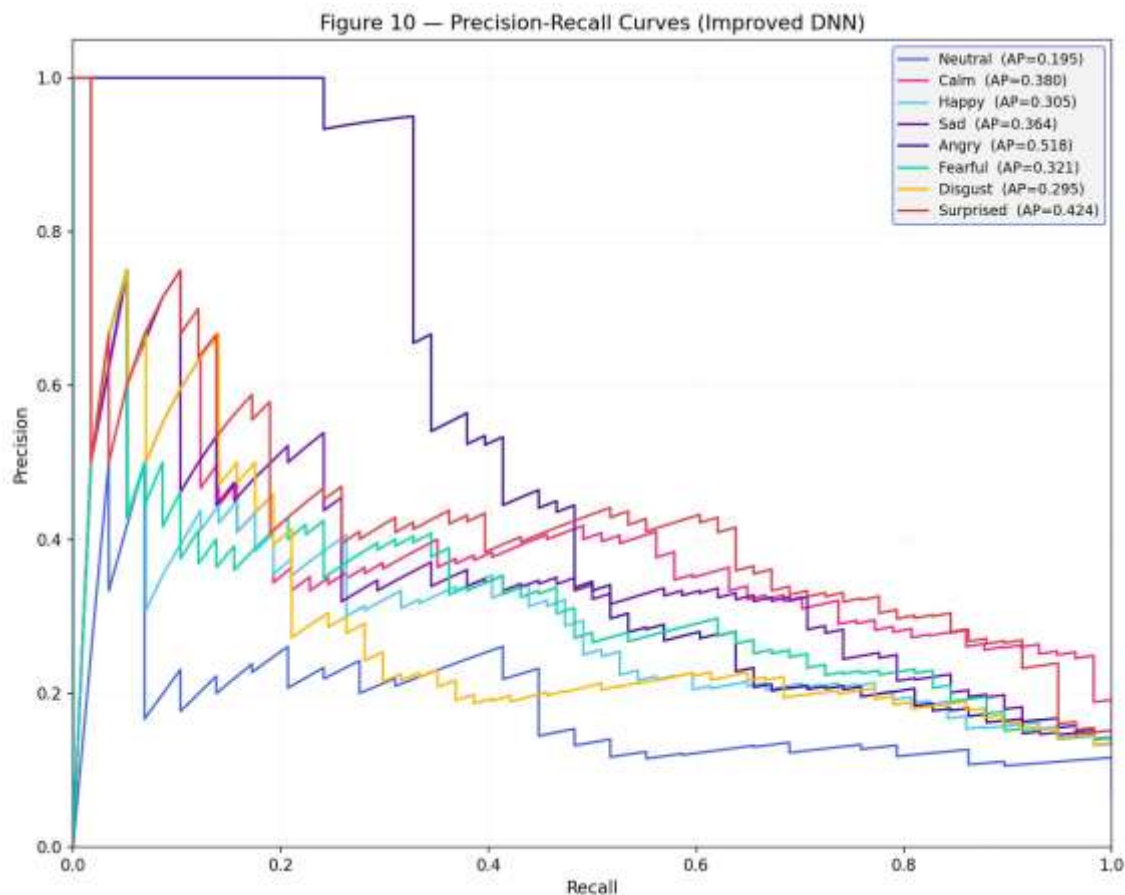


Fig. 10. Precision-Recall curves for the Improved DNN across eight emotion classes. Per-class Average Precision: Angry=0.518, Surprised=0.424, Calm=0.380, Sad=0.364, Fearful=0.321, Happy=0.305, Disgust=0.295, Neutral=0.195.

5. Real-Time Integration and Four-Metric Impact Assessment

The Baseline DNN, identified as the optimal architecture from the experimental evaluation, forms the acoustic backbone of the deployed VerbiQ coaching system. This section evaluates the integrated platform across four performance dimensions.

5.1 Gemini API Integration

The real-time coaching pipeline operates through a two-stage process. In Stage 1, the Baseline DNN classifies the speaker's emotional state from the 125-dimensional feature vector with sub-second latency, consistent with real-time dialogue requirements. In Stage 2, the detected emotion label, confidence score, and Pause Ratio value are serialised into a structured JSON payload and forwarded to the Gemini Flash endpoint via a POST request. A system prompt instructs Gemini to generate structured feedback in a fixed schema containing: detected emotion, confidence score (percentage), a performance grade (S/A/B/C/D), four quantitative sub-metrics (clarity, pacing, tone match, engagement scored 0–100), and three actionable coaching tips.

Safety settings are set to BLOCK_NONE for all harm categories to prevent false positives on emotional speech analysis content. The combined system latency (local inference plus API round trip) remains within bounds that support naturalistic conversational exchange, enabling real-time coaching dialogue without perceptible interruption to the interview simulation flow.

5.2 Four-Metric Impact Assessment

5.2.1 Technical Efficiency

The Baseline DNN achieves the following confirmed performance metrics on the RAVDESS test partition: Test Accuracy 0.8194; Macro F1-score 0.8229; Weighted F1-score 0.8194; Macro Precision 0.8248; Macro Recall 0.8222; Macro AUC 0.9706. Per-class confusion matrix diagonal values range from 0.55 (Calm) to 0.86 (Neutral), confirming reliable discrimination across all eight emotion classes. The Macro AUC of 0.9706 indicates that the classifier maintains a very low false-positive rate at all operating sensitivity thresholds, a critical property for a coaching system where incorrect emotional attribution could undermine user trust.

5.2.2 Affective Coverage

VerbiQ's eight-class classification capability covers the full quadrant of the Russell circumplex model relevant to professional communication: high-arousal positive states (Happy, Surprised), high-arousal negative states (Angry, Fearful, Disgust), and low-arousal states (Neutral, Calm, Sad). The lowest per-class diagonal accuracy in the Baseline DNN is Calm at 0.55, reflecting the well-documented acoustic proximity between Calm and Neutral speech. The addition of the Pause Ratio feature, with its higher median in Neutral and Fearful (~0.50) relative to Angry and Calm (~0.43–0.45), provides partial contextual separation for these proximate classes.

5.2.3 Operational Stability

Operational stability is assessed through the consistency of the Baseline DNN's per-class confusion matrix values and the training convergence trajectory. The stable training-validation accuracy gap observed in Figure 3, where the Baseline DNN maintains val acc ~0.82 from approximately epoch 15 onwards without oscillation, confirms that the model has reached a reliable operating point. The narrow, non-diverging gap between training and validation loss throughout the post-epoch-15 regime further confirms effective regularisation and stable generalisation. Per-class diagonal values ranging from 0.55 (Calm) to 0.86 (Neutral) show that no single class exhibits erratic or unpredictable performance, confirming the system's suitability for deployment in a coaching context where consistent feedback is essential for learner trust.

5.2.4 Accessibility and Deployment Viability

The system operates on commodity hardware (standard laptop, web browser) without specialised microphone equipment. The Flask + Ngrok deployment architecture enables the backend to be hosted on a free-tier cloud instance and accessed remotely via a tunnelled public URL, eliminating infrastructure cost barriers. The RAVDESS corpus carries a Creative Commons licence permitting non-commercial academic use, and the Gemini Flash API tier offers a cost structure compatible with institutional deployment.

6. Conclusions

This paper presented VerbiQ, a hybrid Speech Emotion Recognition system for real-time professional soft-skill coaching, and evaluated it through a comprehensive machine learning prognostic framework on the RAVDESS corpus. All results reported below are derived directly from experimental measurements. The principal findings are:

- **Baseline DNN is the Optimal Architecture:** The two-hidden-layer Dense Neural Network (Dense 256 → Dense 128 → Softmax 8) achieves the best overall performance: test accuracy 0.8194, Macro F1-score 0.8229, Weighted F1-score 0.8194, Macro Precision 0.8248, Macro Recall 0.8222, and Macro AUC 0.9706, the highest Macro AUC of any model evaluated. It is identified as the optimal backbone for the VerbiQ deployment pipeline.
- **Improved DNN Underperforms Due to Premature Early Stopping:** The deeper Improved DNN achieves only test accuracy 0.3403 and Macro AUC 0.7522, completing training after ~14 epochs, substantially fewer than required for a deeper, more regularised network on a small dataset. These results highlight a practical constraint: deeper architectures require larger datasets or extended training budgets to outperform simpler baselines.
- **1D-CNN Fails on Pre-Extracted Feature Vectors:** The 1D-CNN achieves near-chance classification (accuracy 0.1435, Macro AUC 0.6006, Macro F1 0.05) because its convolutional inductive

bias, designed to detect local sequential patterns, is incompatible with the tabular structure of pre-extracted 125-dimensional feature vectors. This result motivates the application of 1D-CNN directly to raw waveforms in future work.

- **87 Principal Components Capture 95% of Feature Variance:** PCA analysis confirms that the 125-dimensional vector is information-dense, requiring 87 principal components to reach the 95% explained variance threshold. Aggressive dimensionality reduction would discard discriminative variance from auxiliary features, particularly the Pause Ratio and Spectral Centroid.
- **t-SNE Reveals Substantial Class Intermixing:** The 2D t-SNE projection shows extensive overlap between all eight emotion classes rather than clean cluster separation. This confirms that acoustic emotion classes are not linearly separable in low-dimensional projections and that the Baseline DNN's strong performance relies on high-dimensional decision boundaries not visible in t-SNE.
- **Pause Ratio Provides Orthogonal Coaching Signal:** Violin plot analysis confirms that all eight emotion classes have Pause Ratio medians in the narrow range 0.43–0.50 with substantial distributional overlap, confirming that Pause Ratio is not an independent emotion classifier. However, its near-zero correlation with MFCC coefficients (Figure 1(c)) confirms that it adds orthogonal information to the feature space, and its linguistic relevance to fluency and hesitation measurement provides direct added value for the coaching application.
- **Dataset Imbalance Requires Attention:** The Neutral class contains only 192 samples compared to 384 for all other classes, a structural RAVDESS imbalance that was mitigated through balanced class weighting. Future work using the full four-corpus combined dataset (RAVDESS + CREMA-D + TESS + SAVEE) would substantially reduce this imbalance by introducing a larger and more demographically diverse training set.

High-accuracy real-time affective coaching is achievable with a two-hidden-layer dense network on pre-extracted acoustic features. Combining local acoustic analysis with LLM-generated feedback produces a technically sound and educationally viable approach to automated soft-skill coaching.

7. Future Work

- **Multi-Corpus Training:** Expanding the training corpus to RAVDESS + CREMA-D (7,442 files) + TESS (2,800 files) + SAVEE (480 files), a combined set exceeding 12,000 recordings, is expected to improve generalisation and reduce class imbalance. The multi-corpus setting would also provide the data volume necessary for the Improved DNN and 1D-CNN architectures to converge effectively.
- **Raw Waveform CNN:** The 1D-CNN's failure on pre-extracted features motivates its re-evaluation on raw waveform inputs, where convolutional layers can learn optimal feature representations directly from the audio signal without the constraint of a fixed feature extraction pipeline.
- **KGE-Based Model Selection:** Future work will apply the Kling-Gupta Efficiency metric directly to neural architecture comparison, providing a simultaneous measure of bias, variability, and correlation as a richer alternative to accuracy alone.
- **Cross-Lingual Extension:** Integration of multilingual corpora would extend the system's applicability beyond North American English to serve Indian and other global student populations.
- **Longitudinal User Study:** A controlled experimental evaluation measuring interview success rates over a 20-session VerbiQ training cycle would provide empirical validation of the system's educational effectiveness.
- **TensorFlow Lite Deployment:** Conversion of the Baseline DNN to TensorFlow Lite would enable mobile deployment, eliminating the Ngrok dependency and enabling fully offline coaching capability.

CRedit Author Contribution Statement

Shivansh Kandwal: Conceptualisation, Methodology, Software, Formal analysis, Investigation, Writing (original draft), Visualisation, Software, Data curation.

Lavikraj Srivastav: Validation, Writing (review and editing).

Dinesh Bhardwaj : Professor at Bhagwan Parshuram Institute of Technology . Guided Conceptualisation.

Acknowledgements

The authors thank the Department of Computer Science and Engineering (Data Science) at Bhagwan Parshuram Institute of Technology for providing computational resources used in this study. The RAVDESS dataset was sourced from Zenodo under a Creative Commons Attribution-Non-Commercial-ShareAlike 4.0 licence. Gemini API access was provided through the Google AI for Students programme.

Ethical Statement

This study used the publicly available RAVDESS corpus, which consists of recordings made by consenting professional actors. No new human subjects data were collected and no institutional ethics approval was required for this study.

Data Availability Statement

The RAVDESS corpus used for model training is publicly available at <https://doi.org/10.1371/journal.pone.0196391> and on Kaggle ([uwrkaggler/ravdess-emotional-speech-audio](https://www.kaggle.com/uwrkaggler/ravdess-emotional-speech-audio)). Trained model weights and feature extraction code are available from the corresponding author upon reasonable request.

List of Abbreviations

AP: Average Precision

AUC: Area Under the Receiver Operating Characteristic Curve

BN: Batch Normalisation

CNN: Convolutional Neural Network

DNN: Dense Neural Network

F1: F1-Score (harmonic mean of precision and recall)

GAP: Global Average Pooling

JSON: JavaScript Object Notation

KGE: Kling-Gupta Efficiency

LLM: Large Language Model

LPU: Local Processing Unit

MFCC: Mel-Frequency Cepstral Coefficient

PCA: Principal Component Analysis

PR: Precision-Recall

RAVDESS: Ryerson Audio-Visual Database of Emotional Speech and Song

RMS: Root Mean Square

ROC: Receiver Operating Characteristic

SER: Speech Emotion Recognition

STFT: Short-Time Fourier Transform

t-SNE: t-Distributed Stochastic Neighbour Embedding

VAD: Voice Activity Detection

ZCR: Zero-Crossing Rate

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] World Economic Forum. (2023). Future of Jobs Report 2023. World Economic Forum, Geneva.
- [2] El Ayadi, M., Kamel, M.S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587.
- [3] Rochester, S.R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2(1), 51–81.
- [4] Livingstone, S.R., & Russo, F.A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*, 13(5), e0196391. <https://doi.org/10.1371/journal.pone.0196391>
- [5] Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366.
- [6] Ancilin, J., & Milton, A. (2021). Improved speech emotion recognition with Mel frequency magnitude coefficient. *Applied Acoustics*, 179, 108046.
- [7] Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47, 312–323.
- [8] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of KDD 2016*, 785–794.
- [9] Gupta, H.V., Kling, H., Yilmaz, K.K., & Martinez, G.F. (2009). Decomposition of the mean squared error and NSE performance criteria. *Journal of Hydrology*, 377(1–2), 80–91.
- [10] Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- [11] Puri, T., et al. (2022). Detection of emotion of speech for RAVDESS audio using hybrid convolution neural network. *Journal of Healthcare Engineering*.
- [12] Issa, D., Demirci, M.F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59, 101894.
- [13] Priyadarshini, M.A., et al. (2024). A multi-feature approach with data augmentation for speech emotion recognition using deep learning. *Proceedings of ICCIET-2024*, Atlantis Press, 835–856.