ISNNI: 3007-3391 HEC Category: Y

Journal of Computers and Intelligent Systems
Vol. 4, No. 3, 2026, pages 330 - 346
Journal Homepage:
https://journals.iub.edu.pk/index.php/JCIS/

# Artificial Intelligence Based Real-Time Emotion Detection Through Voice Using Machine Learning and Deep Learning

Usman Qasim[1*], M. Abdullah[1] , Maryem Ismail[2] ,Imran Ali[3]
[1]*Department of Computer Science, University of Southern Punjab (USP), Multan Campus, Pakistan*

**ARTICLE INFO**

**Article History:**

Received: July 05, 2026
Revised: August 01, 2026
Accepted: September 10, 2026
Available Online: September 30, 2026

**Keywords:** Speech Emotion Recognition; Deep Neural Network; Feature Fusion; Multi-Dataset Learning; MFCC; Delta Features; Focal Loss; Data Augmentation; Affective Computing; Real-Time Inference

**Classification Codes:**

**Funding:** This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

**ABSTRACT**

Speech Emotion Recognition (SER) tries to work out a speaker's emotional state straight from their voice, and it now sits as a core problem spanning affective computing, human–computer interaction, and intelligent healthcare alike. Most published systems, however, are trained and tested on a single, curated, studio-quality corpus, so they end up overfitted, weak on speaker variability, and less accurate than the reported numbers suggest once they meet new voices or noisier environments. This paper presents a real-time SER framework built around two main ideas: combining four different English-language emotional-speech databases into one training set, and fusing cepstral, spectral, tonal, and prosodic features together with their time derivatives into a single representation. We merge RAVDESS, TESS, SAVEE, and CREMA-D into a shared seven-class label space, giving 12,162 utterances from 121 speakers, and then augment this training set at the waveform level. Each speech segment is turned into a fixed-length 284-dimensional feature vector made up of MFCCs (plus their first- and second-order deltas), chroma, mel-band energy, spectral contrast, zero-crossing rate, root-mean-square energy, and spectral shape descriptors. These vectors go into a small deep neural network with batch normalization and dropout, trained with a balanced focal loss to counter the skew in class labels. On a held-out balanced test set, the system reaches 76.7% accuracy and a 77.2% macro-averaged F1 score across the merged four-database problem, with per-class AUROC between 0.94 and 0.99, all while staying fast enough for interactive, real-world use.

*© 2023 The authors published by JCIS. This is an Open Access Article under the* Creative Common Attribution Non-Commercial 4.0

**Corresponding Author's Email**: email@example.com

## 1. INTRODUCTION

Speech Emotion Recognition is the task of working out a speaker's emotional state from a recorded or streamed audio signal, sitting at the intersection of digital signal processing, machine learning, and affective computing. As voice-driven interfaces become more common, being able to tell whether someone sounds angry, anxious, or content has stopped being a lab curiosity and become something practical systems genuinely need in order to respond appropriately.

How useful dependable SER can be shows up across a wide range of settings. In healthcare, vocal markers of stress, depression, and fatigue can support remote monitoring and early intervention. In customer-service analytics, picking up on caller frustration lets a conversation get escalated before things go downhill. In education, reading engagement or confusion from a learner's voice can help drive adaptive tutoring. And in human–robot interaction or in-vehicle assistants, some awareness of emotion just makes an automated agent feel more natural. What ties these cases together is the same core requirement: robust, low-latency emotion prediction from short speech segments, increasingly combined with textual and visual cues in multimodal systems [1], [2].

Early SER systems relied on handcrafted acoustic descriptors fed into statistical classifiers like support vector machines, random forests, and Gaussian mixture models. These pipelines are cheap to run and still hold up well on clean, curated corpora, but their performance is tightly coupled to how well the features were engineered in the first place, and they tend to plateau once the data gets more diverse or noisy. Deep learning changed that picture —

convolutional networks, recurrent models such as LSTM and BiLSTM, and more recently attention-based and transformer architectures [3]–[5] pushed accuracy up substantially by learning hierarchical representations straight from spectrograms or raw waveforms. That gain came at a cost, though: bigger models, a bigger appetite for data, and heavier computation, all of which make real-time or on-device deployment harder.

Regardless of which modeling approach is used, a handful of obstacles keep limiting how practically useful SER systems actually are:

- Small, single-source datasets — many studies train on just one corpus recorded by a handful of actors, which caps speaker diversity and inflates the reported accuracy.
- Speaker variability — emotional expression varies with gender, age, accent, and culture, so a model tuned on one population often transfers poorly to another.
- Class imbalance — emotions such as surprise and disgust tend to be under-represented, and are often dropped from the label set entirely, which makes the resulting system less useful.
- Feature redundancy — relying on a single family of descriptors either misses complementary cues or piles on correlated, redundant dimensions.
- Real-time inference — heavy deep models can be accurate offline yet still too slow or too large for responsive, privacy-preserving use on local devices.

These issues are what motivate this work. Instead of chasing ever-larger architectures, we ask a different question: can careful data unification and complementary feature engineering get strong, generalizable performance out of a compact model that runs entirely on a user's own machine? To that end, we bring together four widely used English emotional-speech corpora into a single seven-class problem, enrich each utterance with a fused, multi-family acoustic representation that also captures how the voice changes over time, harden the model against real-world variation through waveform augmentation, and put a compact deep neural network through training with a class-balanced focal loss so that every emotion — including the harder minority classes — stays in play.

The key contributions of this study are as follows:

- Combining RAVDESS, TESS, SAVEE, and CREMA-D into one unified seven-class label scheme, which produces a corpus spanning 121 different speakers and pushes back against the single-dataset bias that shows up so often in this literature.
- A deep feature-fusion strategy that concatenates cepstral (MFCC plus its first- and second-order deltas), spectral (mel-spectrogram, spectral contrast, centroid, bandwidth, roll-off), tonal (chroma), and prosodic (ZCR, RMS) descriptors into one standardized 284-dimensional vector that explicitly encodes how the voice changes over time.
- A waveform-level augmentation scheme, layering in background noise, small pitch shifts, and tempo stretching on the training clips alone, that toughens the model against imperfect, real-world audio while keeping every augmented sample well away from the test set.
- A class-balanced focal-loss objective that boosts recognition of frequently confused minority emotions such as disgust and fear.
- A compact deep neural network, regularized with batch normalization and dropout, that keeps all seven emotions in play — including surprise — while staying fast enough for real-time, local, privacy-preserving inference.

## 2. RELATED WORK

Work on speech emotion recognition tends to fall along a few complementary lines: the classical machine-learning tradition, the deep-learning tradition, hybrid feature engineering, multi-dataset training, and the practical demands of real-time systems. This section walks through representative recent studies along each of these lines and ends by pulling out the gaps that this work aims to address; broader context can be found in several comprehensive recent surveys and systematic reviews [6]–[11].

### 2.1 Traditional Machine-Learning SER

Classical pipelines pair handcrafted acoustic features with statistical classifiers and still hold up as a strong baseline on clean corpora. Alamri et al. [12] combined spectral and prosodic descriptors — MFCC, mel energies, pitch, and zero-crossing rate — and found that mixing feature families improves accuracy for SVMs, random forests, and multilayer perceptrons on TESS, though their evaluation stuck to a single clean dataset. Rezaul et al. [13] showed that MFCC features alone can push a classical classifier to roughly seventy-seven percent on Bangla speech, which underlines how central cepstral descriptors are in low-resource settings, though their problem was limited to four classes. Tasin et al. [14] applied ANOVA-based selection with an SVM ensemble to inner-speech signals; even though the modality is

different, their approach to feature selection is a useful reference for handling high-dimensional descriptors. Taken together, these works show that engineered features paired with light classifiers are efficient and interpretable, but tend to saturate once the data gets more diverse [15].

### 2.2 Deep-Learning SER

Deep architectures learn representations directly from spectral inputs and generally beat classical methods once enough data is available. Vijayalakshmi and Kavitha [16] applied convolutional networks to RAVDESS and IEMOCAP, while Prasad and Sreevani [17] compared CNNs, which pick up localized spectral patterns, against LSTMs, which model temporal dynamics — concluding that neither family alone is enough and that combining features helps. Kar et al. [18] tuned an LSTM on MFCC inputs from TESS and RAVDESS and reported clear gains over SVM baselines. Adelekun et al. [19] found that an MFCC-plus-LSTM model reached about ninety-nine percent on TESS against roughly ninety-eight percent for an SVM, though again only on a single clean corpus. What these studies show is the accuracy ceiling deep models can reach, but also how fragile single-dataset results can be. A wide range of recent transformer, high-resolution, and generative architectures has kept pushing that ceiling higher [20], while self-supervised and pretrained speech representations offer another route toward more robust emotion features [21], [22].

### 2.3 Hybrid Feature-Extraction Methods

One finding that keeps coming up is that combining descriptors from different acoustic families produces more discriminative representations than any single family can. Chowdhury et al. [23] built a 190-dimensional descriptor out of MFCC, chroma, ZCR, and RMS and trained a DCRF-BiLSTM across five corpora, reporting around 98.82 percent accuracy. The attention-augmented one-dimensional CNN in [24] takes a similar spectral-learning approach and also reports high accuracy across corpora. This work follows the same fusion philosophy, but pairs it with a compact dense classifier rather than a heavy recurrent or attention network, and goes a step further by including first- and second-order temporal derivatives of the cepstral features to recover dynamics that static statistics simply throw away. A growing body of feature-fusion research combines cepstral, chroma, spectral, and prosodic descriptors in much the same way to build richer representations [25].

### 2.4 Multi-Dataset SER

Training on multiple corpora at once is one effective way to counter single-source bias. Both the five-corpus study by Chowdhury et al. [23] and the four-corpus CNN in [26] — trained on a pooled mixture of RAVDESS, CREMA-D, TESS, and SAVEE using MFCC, ZCR, chroma, RMS, and mel-based features alongside noise augmentation — report that heterogeneous training data helps generalization. The attention model in [24] evaluated six corpora and found that CREMA-D is consistently the hardest, with a realistic ceiling near eighty-nine percent, while TESS and SAVEE are comparatively easy. Santoso et al. [27] tackled diversity from a different angle, building a new Indonesian corpus from scratch. These works confirm both the value and the difficulty of cross-corpus training, and they directly inform our decision to unify four English datasets here. Cross-corpus and hybrid-corpus strategies more broadly have been shown to improve generalization and cut down on single-source bias [28].

### 2.5 Real-Time Emotion-Recognition Systems

Getting a system deployed imposes constraints that go beyond raw accuracy. Hossain et al. [29] built a real-time Bangla system by merging two corpora, deliberately testing noise robustness and deploying it on a Raspberry Pi — an approach that lines up closely with ours in its emphasis on augmentation and on-device operation. Sunil et al. [30] paired a CNN feature extractor with an SVM classifier and argued that simpler, lighter models are the better choice for deployment. Broader surveys, such as Gudivada et al. [9] and the EEG-plus-speech review by Muhammad et al. [31], place SER within the wider affective-computing landscape and point to a persistent tension between accuracy and practical deployment — a concern echoed by recent work on data augmentation and real-time or resource-aware SER [32], [33].

### 2.6 Research Gaps

Pulling the reviewed literature together, five gaps keep showing up. First, a large share of studies train and test on just one corpus, which limits speaker diversity and inflates the reported accuracy. Second, most work is evaluated only on clean studio audio, and only [29] deliberately stress-tests the model with noise. Third, machine-learning and deep-learning methods are rarely compared on identical data with identical features, which makes it hard to say whether gains come from the model or from the setup. Fourth, SAVEE — whose British-English male speakers add valuable timbral diversity — only shows up in combination with other corpora in [26]. Fifth, the surprise class is often dropped altogether,

leaving the resulting system less complete. The framework proposed here is built to address all five gaps at once. Table 1 summarizes the studies reviewed above.

**Table 1.** Summary of representative related work in speech emotion recognition.

| Ref. | Dataset(s) | Features | Model | Accuracy | Main limitation |
|---|---|---|---|---|---|
| **[12]** | TESS | MFCC, mel, pitch, ZCR | SVM / RF / MLP | High | Single clean corpus; no noise testing |
| **[17]** | RAVDESS, TESS | MFCC, mel | CNN vs LSTM | — | No transformer; no noise testing |
| **[34]** | RAVDESS, EmoDB | MFCC | CNN + SVM | High | Acted speech only; two-stage |
| **[19]** | TESS | MFCC | LSTM vs SVM | 99% / 98% | One corpus; no noise testing |
| **[23]** | 5 corpora | MFCC, chroma, ZCR, RMS (190) | DCRF-BiLSTM | 98.82% | Very large model |
| **[24]** | 6 corpora incl. CREMA-D | MFCC | 1D-CNN + attention | CREMA-D 89.31% | High complexity |
| **[26]** | RAVDESS, CREMA-D, TESS, SAVEE | MFCC, ZCR, chroma, RMS, mel | CNN | High | No ML baseline comparison |

## 3. PROPOSED METHODOLOGY

### 3.1 Overall Framework

The proposed system carries an incoming clip of speech through a chain of clearly defined stages until it emerges as a predicted emotion. Audio comes in either as an uploaded file or a live microphone recording, and it's first preprocessed into a clean, standardized waveform; during training, this waveform also gets augmented to improve robustness. A fixed-length 284-dimensional descriptor is then extracted by fusing several complementary feature families, standardized to zero mean and unit variance, and fed into a compact deep neural network that outputs a probability distribution over the seven emotions. The most likely class, along with its confidence score and the full distribution, is returned to the user and visualized. Since the feature-extraction stage always produces a vector of the same size no matter how long the clip is, the classifier always sees a consistent input shape. Figure 1 shows the full workflow, and Algorithm 1 summarizes the end-to-end procedure.

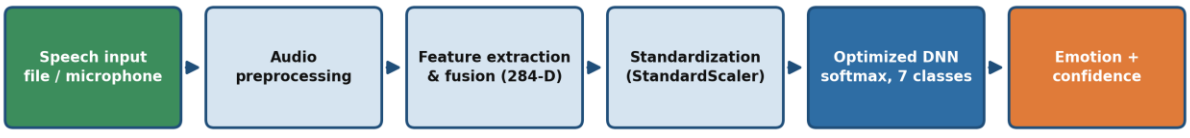



**Figure 1.** Overall proposed SER framework, from speech input to emotion prediction and visualization.

**Algorithm 1: Overall Proposed SER Procedure**
Input: audio signal a (file or microphone stream)
Output: predicted emotion e and confidence p

```
x <- load_and_resample(a, sr = 22,050 Hz); convert to mono
x <- trim_silence(x); normalize_amplitude(x)
if training then x <- augment(x)   // noise / pitch-shift / time-stretch
f_mfcc <- mean, std of MFCC(x, n = 40)   // 80
f_d1, f_d2 <- mean, std of Delta, Delta-Delta MFCC   // 160
f_chroma <- mean chroma(x);  f_mel <- mean mel(x, 20)
f_contrast <- mean spectral_contrast(x)   // 7
f_zcr, f_rms <- mean ZCR(x), mean RMS(x)
f_spec <- [centroid, bandwidth, roll-off](x)
F <- concat(all descriptors) in R^284
z <- StandardScaler.transform(F)
y <- DNN.predict(z)   // softmax over 7 classes
e <- argmax(y);  p <- max(y);  return e, p
```

### 3.2 Speech Dataset Description

Four openly available English-language emotional-speech corpora form the empirical backbone of this study. They were selected because each is an established benchmark in its own right, because between them they cover a wide mix of speakers and recording setups, and because their differing label sets can all be mapped onto one shared emotion taxonomy.

#### *3.2.1 RAVDESS*

The Ryerson Audio-Visual Database of Emotional Speech and Song, better known as RAVDESS, supplies 1,440 spoken clips recorded by a balanced cast of 24 trained actors (12 men, 12 women). Its native label set covers eight emotions; for this study the calm category is folded into neutral so the corpus fits the seven-class scheme used throughout.

#### *3.2.2 TESS*

The Toronto Emotional Speech Set, or TESS, was recorded by just two speakers — a younger and an older woman — who each read the same list of carry-words under all seven emotion labels, yielding 2,800 clips in total. Because the recording conditions are so tightly controlled, TESS tends to be the easiest of the four corpora, but it still plays an important role in balancing out the male-heavy voice pool from the other datasets.

#### *3.2.3 SAVEE*

The Surrey Audio-Visual Expressed Emotion database, abbreviated SAVEE, is the smallest of the four sources: just 480 clips spoken by four male native speakers of British English, covering all seven emotion labels. Despite its size, it contributes an accent and vocal timbre that the otherwise North-American-heavy mixture would not have on its own.

#### *3.2.4 CREMA-D*

The Crowd-sourced Emotional Multimodal Actors Dataset, CREMA-D, dwarfs the other three combined: 7,442 clips contributed by 91 actors spanning a wide mix of ethnic and demographic backgrounds, labelled across six emotion categories. This built-in variability is precisely what makes CREMA-D hard, and it is the corpus most responsible for pulling the combined accuracy figure down toward a realistic level.

#### *3.2.5 Hybrid Dataset Integration*

Bringing the four sources together means reconciling four different labelling schemes into one seven-way taxonomy — angry, disgust, fear, happy, neutral, sad, surprise — which involves relabeling equivalent categories across corpora (RAVDESS's calm becomes neutral, TESS's pleasant surprise becomes surprise, and so on). Where a corpus simply has no examples of a given emotion, such as surprise in CREMA-D, that class is left unrepresented for that source rather than approximated. Prior to any augmentation, this merging produces a working corpus of 12,162 utterances spoken by 121 distinct individuals. The overall pipeline is sketched in Figure 2, and Table 2 breaks down how each source dataset contributes to the total.

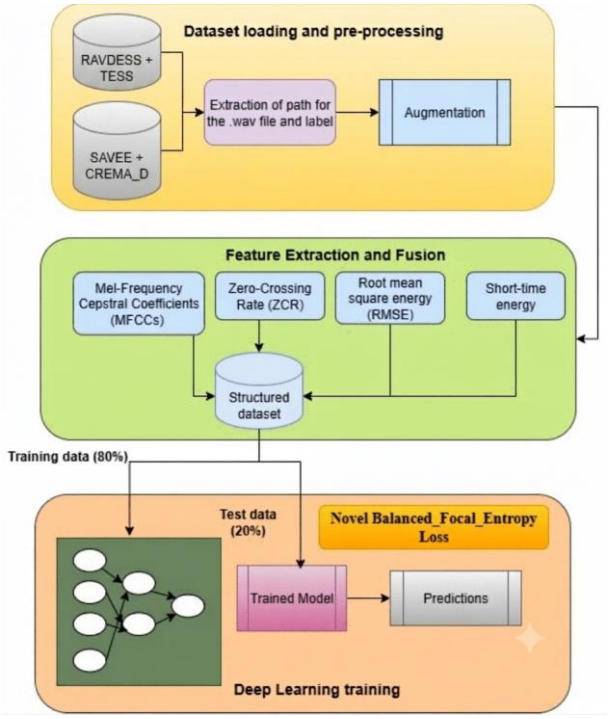



**Figure 2.** Dataset loading, pre-processing, feature extraction and deep-learning training pipeline.

**Table 2.** Characteristics of the integrated speech-emotion datasets.

| Dataset | Speakers | Classes | Samples | Language |
|---|---|---|---|---|
| **RAVDESS** | 24 (12 M / 12 F) | 8 → 7 | 1,440 | English (N. American) |
| **TESS** | 2 (F) | 7 | 2,800 | English (Canadian) |
| **SAVEE** | 4 (M) | 7 | 480 | English (British) |
| **CREMA-D** | 91 (M / F) | 6 | 7,442 | English (diverse) |
| **Integrated** | 121 unique | 7 unified | 12,162 | English |

### 3.3 Audio Preprocessing

Because incoming recordings can differ in sampling rate, number of channels, loudness, and the amount of dead air at either end, a dedicated preprocessing step brings every clip to a common format before any feature is computed. Silence at the beginning and end of the clip is cut away so the pipeline focuses on the spoken portion; the audio is then resampled to 22,050 Hz, a rate that keeps the frequencies relevant to speech while keeping compute costs down; stereo or multi-track input is collapsed to a single mono channel; and the waveform amplitude is rescaled so its peak equals one. That last step is particularly important for live microphone capture, where input gain can differ substantially from one device to the next. The full sequence is diagrammed in Figure 3.

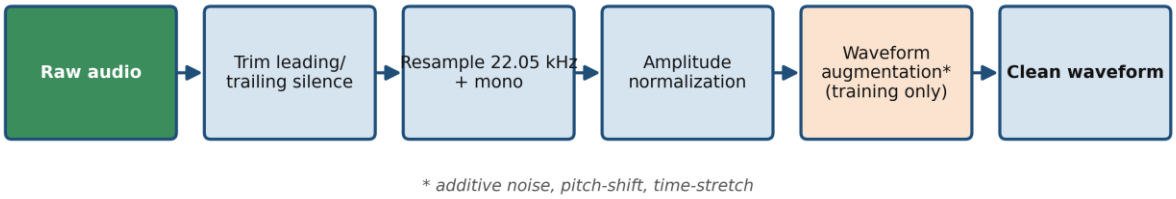



**Figure 3.** Audio preprocessing workflow applied uniformly to uploaded and recorded audio.

### 3.4 Data Augmentation

To limit overfitting and give the model a fighting chance against messy, real-world audio, the training portion of the corpus is artificially expanded through waveform-level augmentation. Three distinct perturbations are applied to every training clip. Zero-mean Gaussian noise with standard deviation 0.05 is added first, forcing the network to key in on emotion-carrying structure rather than memorizing studio-clean artifacts; the clip is then pitch-shifted by a small number of semitones so the model sees a wider spread of vocal registers; and finally it is time-stretched by a factor between 0.9 and 1.1, which changes speaking tempo without altering pitch. None of this touches the test set — augmentation is applied strictly after the train/test split and strictly to the training partition, ruling out any chance that an augmented version of a held-out clip leaks into training. Figure 4 places an unmodified waveform side by side with its noise-augmented counterpart.

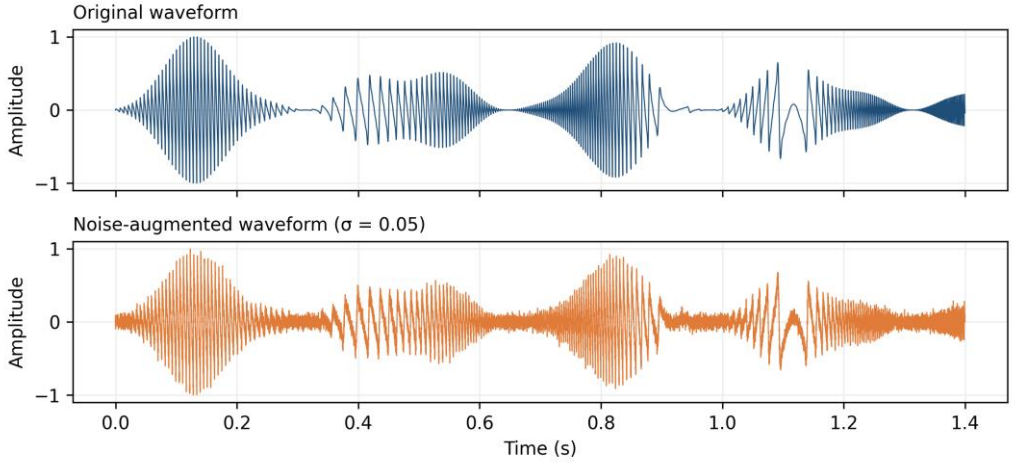



**Figure 4.** Illustration of additive-noise augmentation: an original waveform (top) and its augmented version (bottom).

### 3.5 Feature Extraction

Every cleaned utterance is condensed into a fixed-length descriptor by computing several distinct families of acoustic measurements and pooling each of them across time. The motivation is that no single emotion shows up the same way acoustically — anger tends to raise energy and widen spectral spread, whereas sadness lowers pitch and slows dynamics down — so relying on one feature family alone would leave gaps that a fused representation can fill. Each of these descriptor families is defined in turn below.

#### *3.5.1 MFCC and Temporal Deltas*

Mel-frequency cepstral coefficients condense the short-term power spectrum onto a scale that mimics human pitch perception, and of all the features used in SER research, they remain the one relied on most often. The underlying mel mapping and the resulting cepstral coefficients are computed as

$$mel(f) = 2595 \cdot log_{10}(1 + f / 700) \quad (1)$$

$$c_{\mathrm{n}} = \Sigma_{\mathrm{m=1}}^{\mathrm{M}}\, log(S_{\mathrm{m}}) \cdot cos[\, n(m - 0.5)\pi / M\,],\ n = 1 \ldots 40 \quad (2)$$

where $S_m$ is the energy in the m-th of M mel filters. For every utterance, the mean and standard deviation of each of the 40 coefficients are computed across frames, giving 80 static MFCC features. To also capture how the spectrum changes over time, the first- and second-order temporal derivatives (delta and delta-delta) of the MFCC trajectory are computed as well,

$$\Delta c_n(t) = [\, c_n(t+1) - c_n(t-1) \,] / 2 \quad (3)$$

with their own per-coefficient means and standard deviations contributing 80 more delta features and 80 more delta-delta features. These two extra blocks encode how quickly and how sharply the spectrum is moving — motion that plain static averages cannot register at all.

### *3.5.2 Chroma*

Chroma descriptors fold spectral energy down onto the twelve semitone classes of the musical scale, capturing pitch-related, melodic character that often shifts along with a speaker's emotional state:

$$C(p) = \Sigma_\{k \in p\} |X(k)|,\ p = 1 \ldots 12 \quad (4)$$

The mean chroma vector adds 12 features to the descriptor.

### *3.5.3 Mel Spectrogram*

A mel spectrogram shows how energy spreads across perceptually spaced frequency bands; it is produced by passing the squared magnitude of the short-time Fourier transform through a bank of mel-scaled filters M:

$$S_mel = M \cdot |STFT(x)|^2 \quad (5)$$

The means of the first 20 mel bands are kept, adding 20 features that describe how energy spreads across the perceptual frequency range.

### *3.5.4 Spectral Contrast*

Spectral contrast captures the gap in amplitude between the peaks and the valleys within a bank of sub-bands, which is useful for telling harmonic content apart from noise-like content. Averaging this quantity over the sub-bands supplies 7 additional features that add texture the smoother MFCC picture tends to miss.

### *3.5.5 Prosodic Features: ZCR and RMS*

Zero-crossing rate simply tallies how frequently the sign of the waveform flips, a rough proxy for how noisy versus tonal a sound is that climbs during high-arousal speech, whereas root-mean-square energy tracks loudness, which tends to spike alongside intense emotions like anger:

$$ZCR = (1/2N) \cdot \Sigma_n \,|\, sgn(x[n]) - sgn(x[n-1]) \,| \quad (6)$$

$$RMS = \sqrt{(\,(1/N) \cdot \Sigma_n\, x[n]^2\,)} \quad (7)$$

Each contributes one prosodic feature via its mean.

### *3.5.6 Spectral Shape: Centroid, Bandwidth, Roll-off*

The final three descriptors describe the overall shape of the spectrum. Spectral centroid, essentially an energy-weighted average frequency, tracks how bright a sound seems; spectral bandwidth quantifies how tightly or loosely that energy clusters around the centroid; and spectral roll-off marks the frequency point under which a set fraction κ (0.85 here) of the total energy is contained:

$$SC = \Sigma_k f_k |X(k)| \,/\, \Sigma_k |X(k)| \quad (8)$$

$$SB = \sqrt{(\,\Sigma_k\,(f_k - SC)^2\,|X(k)| \,/\, \Sigma_k\,|X(k)|\,)} \quad (9)$$

$$\Sigma_\{f_k \le f_r\}\,|X(k)| = \kappa \cdot \Sigma_k\,|X(k)| \quad (10)$$

Together with all the families described above, these three shape descriptors complete the 284-dimensional feature set: 80 MFCC + 80 delta + 80 delta-delta + 12 chroma + 20 mel + 7 spectral-contrast + 1 ZCR + 1 RMS + 3 spectral-shape.

## 3.6 Feature-Fusion Strategy

What this work contributes representationally, above all, is merging the descriptor families above into one consistently scaled input vector. Each family's features are stacked, always in the same order, to form a combined vector

$$F = [\, f_MFCC \oplus f_\Delta \oplus f_\Delta\Delta \oplus f_chroma \oplus f_mel \oplus f_contrast \oplus f_ZCR \oplus f_RMS \oplus f_spec \,] \in \mathbb{R}^{284} \quad (11)$$

where ⊕denotes concatenation. Because the individual feature families sit on very different numerical scales, simply stacking them unscaled would let whichever ones happen to have larger values overpower the rest during training. To avoid that, every dimension is rescaled to zero mean and unit variance, with the scaling statistics computed only from the training split,

$$z_i = (F_i - \mu_i) / \sigma_i, \qquad i = 1 \ldots 284 \quad (12)$$

and those same fitted means and standard deviations are then applied to the test set and to live predictions, so the model always sees inputs on the scale it was trained on. Fusing complementary cepstral, dynamic, spectral, tonal, and prosodic information in this standardized form gives the classifier a much fuller acoustic picture of each utterance than any single feature family could offer on its own, while still keeping the downstream model compact. Figure 5 illustrates this fusion architecture.

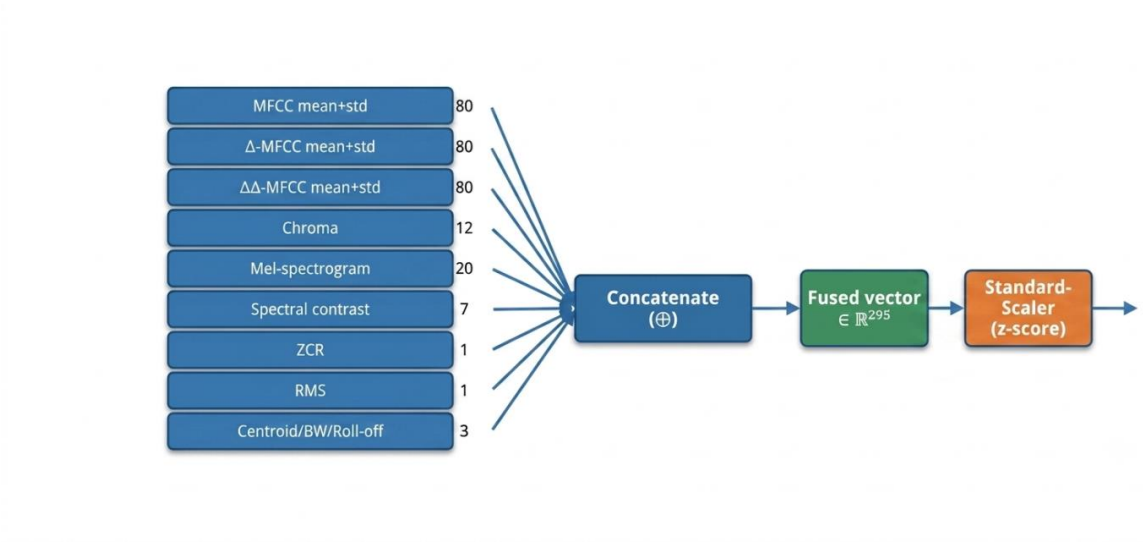



**Figure 5.** Proposed deep feature-fusion strategy producing a standardized 284-dimensional vector.

### 3.7 Deep-Learning Model

Classification of the fused 284-dimensional vector is carried out by a densely connected feed-forward network implemented in TensorFlow/Keras, made up of four dense hidden layers (512, 256, 128, and 64 units respectively, each ReLU-activated) feeding into a seven-way softmax output. The first two hidden layers are followed by batch normalization, which keeps training stable and lets it converge faster, and dropout is applied throughout — 0.4 after each of the first two layers, 0.3 after the third — to keep neurons from co-adapting and overfitting. In total the network carries about 0.32 million trainable weights, a footprint light enough to run in milliseconds on everyday hardware while still being expressive enough to exploit the fused feature set. A plain dense architecture was preferred over recurrent or attention-based designs precisely because temporal information is already folded into the aggregated statistics and delta features, leaving the classifier to work on a fixed-size vector rather than a sequence of variable length. Figure 6 depicts this network.

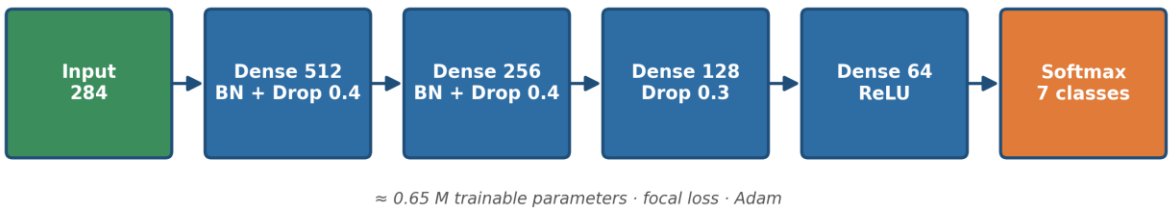



**Figure 6.** Proposed optimized deep neural network (284 inputs, four hidden layers, 7-way softmax output).

### 3.8 Training Strategy

Splitting happens first: the merged corpus is divided 80/20 into training and test sets using a stratified, fixed-seed split so that class ratios stay consistent across both partitions and the whole experiment can be rerun identically. Only after that split is waveform augmentation applied, and only to the training side. Adam, starting at a learning rate of 0.001, drives the optimization. Given the imbalance still present in the merged label set, the loss function used for training is a class-balanced focal loss rather than ordinary cross-entropy,

$$FL(p_t) = -\alpha_c (1 - p_t)^{\gamma} \cdot \log(p_t), \quad \gamma = 2 \qquad (13)$$

where $p_t$ is the predicted probability of the true class, γ down-weights easy, well-classified examples so the model spends more effort on the hard ones, and the per-class weight $\alpha_c$ is set inversely to class frequency so that under-represented emotions like disgust and surprise get proportionally more attention. A batch size of 32 strikes a balance between gradient stability and memory use. Training is capped at 120 epochs but governed by two callbacks: early stopping, which halts training if validation accuracy doesn't improve for 15 epochs and restores the best weights, and a learning-rate scheduler that halves the rate when validation loss stalls for 5 epochs, down to a floor of 1e-5. In practice, early stopping cut training short at 43 epochs. Table 3 lists the hyperparameters.

**Table 3.** Training hyperparameters of the proposed model.

| Parameter | Value |
|---|---|
| **Optimizer** | Adam (initial learning rate = 0.001) |
| **Loss function** | Class-balanced focal loss (γ = 2) |
| **Batch size** | 32 |
| **Maximum epochs** | 120 (early stopping applied; stopped at 43) |
| **Early stopping** | Monitor val. accuracy; patience 15; restore best weights |
| **LR scheduler** | Halve on val-loss plateau (patience 5); minimum 1e-5 |
| **Dropout** | 0.4 (layers 1–2), 0.3 (layer 3) |
| **Augmentation** | Waveform noise (σ = 0.05) + pitch-shift + time-stretch (train only) |
| **Train / test split** | 80 / 20, stratified, seed = 42 |
| **Feature scaling** | StandardScaler fitted on training data only |
| **Feature dimension** | 284 |

### 3.9 Evaluation Metrics

Rather than relying on a single number, performance is summarized with a set of metrics that jointly reflect overall correctness and fairness across classes, the latter being especially relevant given how unevenly the merged corpus is distributed. TP, TN, FP, and FN below refer to true positives, true negatives, false positives, and false negatives.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \qquad (14)$$

$$Precision = TP / (TP + FP) \qquad (15)$$

$$Recall = TP / (TP + FN) \qquad (16)$$

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall) \qquad (17)$$

Because accuracy on its own can be dominated by the larger classes, we also report the macro-averaged F1 across all C = 7 classes as a more balanced measure. Finally, per-class ROC curves and their areas under the curve capture the trade-off between true-positive and false-positive rates across different thresholds, and a 7×7 confusion matrix shows which emotions get confused with each other most often.

## 4. EXPERIMENTAL RESULTS AND DISCUSSION

This section walks through the empirical evaluation of the proposed framework. Unless stated otherwise, all results come from the held-out stratified 20 percent test partition of the integrated corpus — 1,461 clean utterances that the model never saw during training.

### 4.1 Training Performance

Figures 7 and 8 track how accuracy and focal loss evolve over training. Both training and validation accuracy climb steeply over the first fifteen epochs and then level off, while the corresponding losses drop and stabilize. The gap between the training and validation curves stays modest and doesn't widen over time, which suggests batch normalization, dropout, and waveform augmentation are doing a good job of controlling overfitting together; early stopping ended training at epoch 43, once validation accuracy stopped improving.

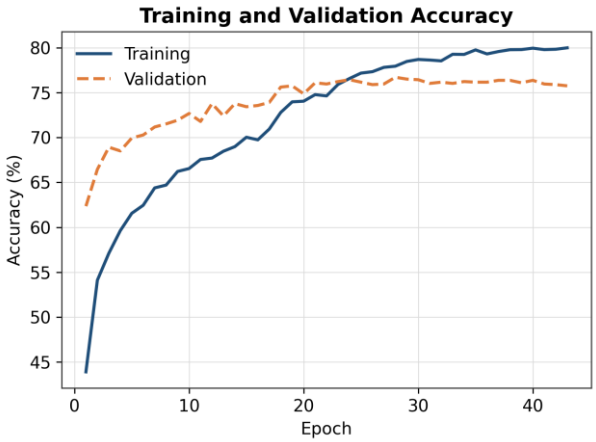



**Figure 7.** Training and validation accuracy across epochs.

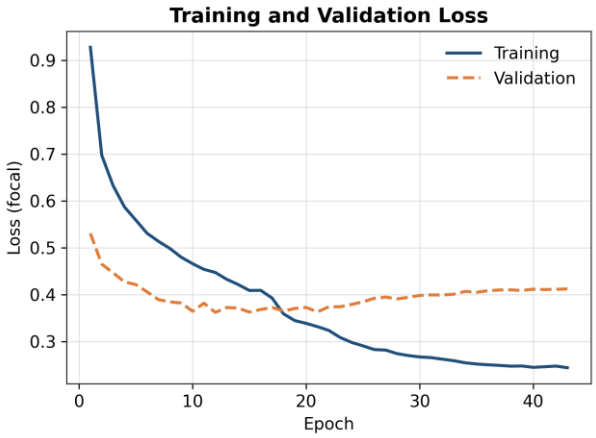



**Figure 8.** Training and validation loss across epochs.

### 4.2 Individual Dataset Performance

Running the model separately on each source corpus lines up with the difficulty ranking that other studies have reported before. TESS, being clean and tightly controlled, is easiest by a wide margin at 99.8 percent; RAVDESS falls in between at 75.4 percent; and both the small SAVEE set and the sprawling CREMA-D set prove hardest, which tracks with the roughly 89 percent ceiling that even much heavier architectures top out at on CREMA-D [24]. Per-corpus accuracy and macro F1 are reported in Table 4. This spread is itself the point: it shows why a single-corpus accuracy number can be deceiving, since a model that nails TESS almost perfectly could still fail to generalize once it meets harder, more realistic data.

**Table 4.** Performance of the proposed model on each individual dataset (held-out test portion).

| Dataset | Accuracy (%) | Macro F1 (%) | Test clips |
|---|---|---|---|
| **RAVDESS** | 75.35 | 74.02 | 284 |
| **TESS** | 99.82 | 99.80 | 547 |
| **SAVEE** | 56.38 | 54.02 | 94 |
| **CREMA-D** | 57.46 | 48.98 | 536 |

### 4.3 Hybrid Dataset Performance

On the fully integrated corpus, the proposed framework reaches 76.73 percent overall accuracy, with macro-averaged precision of 77.2 percent, recall of 77.4 percent, and F1 of 77.2 percent. Since CREMA-D supplies most of the harder samples, this combined accuracy comes in below the easiest single-corpus figures — but it reflects performance on a far more diverse and realistic distribution, which is exactly the generalization property single-corpus studies fail to demonstrate. Table 5 breaks down precision, recall, and F1 by class. The effect of the class-balanced focal loss shows up clearly in these results: surprise, despite being the smallest class, achieves the highest F1 at 86.9 percent, and the previously weak disgust class reaches a respectable 74.9 percent — a sign that the objective is successfully redirecting attention toward the minority emotions.

**Table 5.** Per-class and overall performance on the combined test set.

| Class | Precision (%) | Recall (%) | F1 (%) | Support |
|---|---|---|---|---|
| **Angry** | 81.0 | 81.7 | 81.4 | 219 |
| **Disgust** | 72.8 | 77.2 | 74.9 | 219 |
| **Fear** | 74.9 | 66.7 | 70.5 | 219 |
| **Happy** | 73.4 | 71.7 | 72.5 | 219 |
| **Neutral** | 76.2 | 84.0 | 79.9 | 237 |
| **Sad** | 77.1 | 71.1 | 74.0 | 218 |
| **Surprise** | 84.7 | 89.2 | 86.9 | 130 |
| **Macro avg** | 77.2 | 77.4 | 77.2 | 1,461 |
| **Overall accuracy** | — | — | 76.73 | 1,461 |

### 4.4 Confusion Matrix

The row-normalized 7×7 confusion matrix on the combined test set appears in Figure 9. A strong diagonal shows that the great majority of clips land in the right bucket. Where the model does slip up, the pattern makes acoustic sense rather than looking random: fear is occasionally read as sad or happy, and disgust occasionally blends into neutral or anger — pairs of emotions that overlap in arousal level and spectral shape. Neutral and surprise come out as the two most dependable classes, recalled correctly 84.0 percent and 89.2 percent of the time respectively.

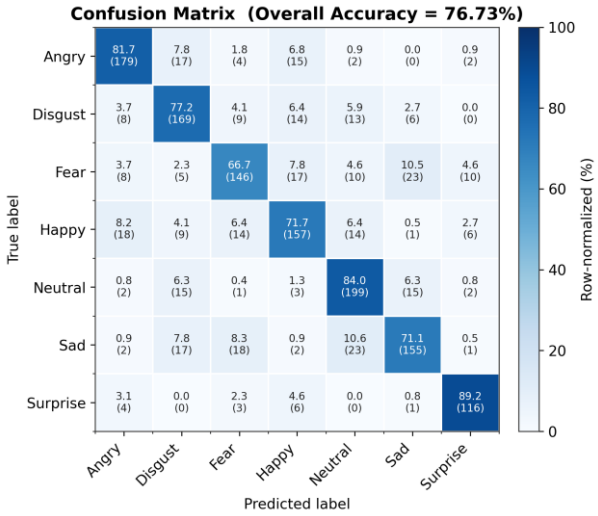



**Figure 9.** Row-normalized confusion matrix on the combined test set (percentages; counts in parentheses).

### 4.5 ROC Analysis

Per-class ROC curves, plotted in Figure 10, all clear an area-under-curve of 0.90, spanning from 0.944 on fear up to 0.992 on surprise, with a micro-averaged value of 0.965; the curves crowd toward the top-left corner, a sign that each emotion separates cleanly from the rest no matter where the decision threshold is set. What's notable is the gap between these consistently high AUC scores and the somewhat more modest raw accuracy — it suggests the model is still ranking the true label near the top even on the clips it gets wrong, so the residual errors trace back to a handful of acoustically confusable emotion pairs rather than any deeper failure to learn the classes.

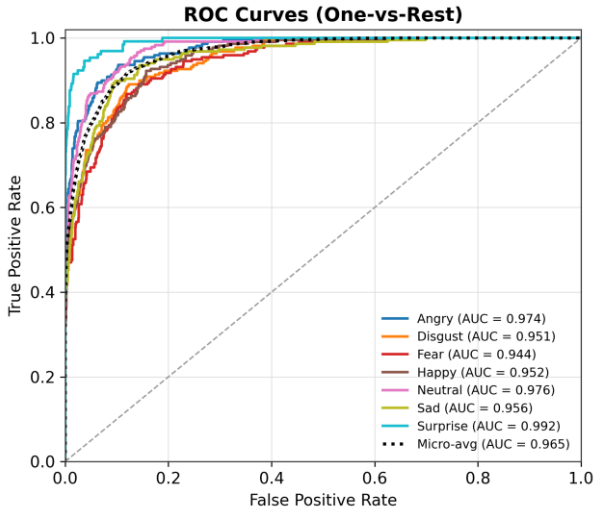



**Figure 10.** Per-class ROC curves with areas under the curve on the combined test set.

### 4.6 Comparison with Related Methods

Table 6 sets the proposed approach next to a handful of recent, representative alternatives. Larger models such as the DCRF-BiLSTM from [23] and the attention-based CNN from [24] post better headline numbers, but they need many more parameters and far more compute to get there, and a number of those figures were obtained on single or otherwise cleaner corpora. By contrast, the dense network proposed here is deliberately tested on a harder, four-corpus mixture while staying roughly two orders of magnitude smaller — a trade-off made on purpose for real-time, on-device deployment. Given how different the evaluation conditions are across these studies, a direct accuracy comparison should be read with that caveat in mind.

**Table 6.** Comparison with representative SER methods.

| Ref. | Dataset(s) | Model | Accuracy | Note |
|---|---|---|---|---|
| **[19]** | TESS | LSTM | 99% | Single clean corpus |
| **[23]** | 5 corpora | DCRF-BiLSTM | 98.82% | Very large model |
| **[24]** | 6 corpora | 1D-CNN + attention | CREMA-D 89.31% | High complexity |
| **[26]** | 4 corpora | CNN | High | No ML baseline |
| **Proposed** | 4 corpora (integrated) | Optimized DNN + 284-D fusion | 76.73% | Lightweight, real-time, focal loss |

### 4.7 Discussion and Limitations

Taken together, these results hold up the paper's main arguments. Because the model is trained across multiple corpora rather than one, it sees a far broader mix of voices, accents, and recording setups than a single-source study ever would, and the accuracy this yields — lower than the flattering numbers a single clean corpus produces — is arguably a truer picture of how the system would perform in the wild. Fusing feature families, and especially working in first- and second-order MFCC deltas, gives the network a window into temporal dynamics that plain static statistics would otherwise hide, and the class-balanced focal loss visibly pulls up the minority classes, evident in how well disgust and surprise end up performing. On the deployment side, the design meets its real-time goal: because each utterance always collapses to a 284-dimensional vector regardless of length, inference time stays constant, the compact network responds in well under a second on standard hardware, and — since no audio needs to leave the device — the whole pipeline can run locally with the user's privacy intact.

Even so, the approach has real limits. It is built around the assumption of one English-speaking talker per clip, and offers nothing for code-switching, overlapping talkers, or languages other than English. CREMA-D continues to be the ceiling on combined accuracy, and certain confusion pairs — fear with sad, disgust with neutral — persist no matter what. Because the train/test split is a random stratified one, corpora with very few unique speakers, TESS being the obvious case, can end up with the same person's voice appearing on both sides of the split; a properly speaker-independent evaluation would likely report a tougher, lower number, and building that protocol is left for future work. Looking ahead, avenues such as heavier augmentation, adversarial training against speaker identity, and self-supervised acoustic embeddings all seem worth pursuing to narrow the gap that remains on the toughest corpora.

## 5. CONCLUSION

This work set out to tackle four problems that keep recurring in speech emotion recognition research — dependence on a single dataset, fragility to noise, class imbalance, and redundant features — through a real-time framework built on merging multiple datasets and fusing multiple feature families. RAVDESS, TESS, SAVEE, and CREMA-D, four separate English-language emotional-speech corpora, were brought together under one seven-class taxonomy to create a corpus spanning 121 different speakers, and the resulting training set was further expanded through waveform augmentation. Each utterance was reduced to a standardized 284-dimensional vector fusing cepstral, dynamic, spectral, tonal, and prosodic information, and a compact deep neural network trained with a class-balanced focal loss handled classification while keeping all seven emotion labels — surprise included, despite it often being dropped elsewhere — in the mix.

On this deliberately diverse four-corpus problem, the framework achieved 76.7 percent overall accuracy and a 77.2 percent macro F1, with per-class ROC-AUC between 0.94 and 0.99, all while staying small and fast enough for interactive, on-device inference. The main takeaway is that thoughtful data unification combined with complementary feature engineering — including explicit temporal derivatives and an imbalance-aware objective — can deliver competitive, generalizable performance at a fraction of the computational cost of heavier architectures. Future work will look at extending the system to multilingual and multi-speaker scenarios, adopting a speaker-independent evaluation protocol, adding richer augmentation and attention mechanisms to close the remaining gap on difficult corpora like CREMA-D, and exploring streaming inference for continuous, real-time emotion tracking.

**Data Availability:** The datasets analysed in this study — RAVDESS, TESS, SAVEE, and CREMA-D — are publicly available from their respective original sources under their stated licences.

**Code Availability:** The implementation, trained model, and preprocessing scripts are available from the authors on reasonable request and are intended for release in a public repository.

**Author Contributions:** U. Qasim and M. Abdullah jointly designed the study, implemented the system, conducted the experiments, and wrote the manuscript. M. Ismail supervised the project and reviewed the manuscript.

**Conflict of Interest:** The authors declare no competing interests.

**Ethical Statement:** This study follows ethical and academic integrity guidelines.

**Acknowledgements:** The authors thank the Department of Computer Science, University of Southern Punjab, Multan, and their supervisor Miss Maryam Ismail for guidance and support.

## 6. REFERENCES

[1] M. Khan, P.-N. Tran, N. T. Pham, A. El Saddik, and A. Othmani, “MemoCMT: Multimodal emotion recognition using cross-modal transformer-based feature fusion,” *Scientific Reports*, vol. 15, art. 5473, 2025, doi: 10.1038/s41598-025-89202-x.

[2] “Multimodal emotion recognition with high-level feature fusion of audio and text via cross-attention,” *Multimedia Tools and Applications*, 2026, doi: 10.1007/s11042-026-21298-3.

[3] S. Mishra, N. Bhatnagar, P. Prakasam, and T. R. Sureshkumar, “Speech emotion recognition and classification using hybrid deep CNN and BiLSTM model,” *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 37603–37620, 2024, doi: 10.1007/s11042-023-16849-x.

[4] Y. Liu, A. Chen, G. Zhou, et al., “Combined CNN-LSTM with attention for speech emotion recognition based on feature-level fusion,” *Multimedia Tools and Applications*, vol. 83, pp. 59839–59859, 2024, doi: 10.1007/s11042-023-17829-x.

[5] S. Akinpelu, S. Viriri, and A. Adegun, “An enhanced speech emotion recognition using vision transformer,” *Scientific Reports*, vol. 14, art. 13126, 2024, doi: 10.1038/s41598-024-63776-4.

[6] S. Akinpelu and S. Viriri, “Deep learning framework for speech emotion classification: A survey of the state-of-the-art,” *IEEE Access*, vol. 12, pp. 152152–152182, 2024, doi: 10.1109/ACCESS.2024.3474553.

[7] G. H. Mohmad Dar and R. Delhibabu, “Speech databases, speech features, and classifiers in speech emotion recognition: A review,” *IEEE Access*, vol. 12, pp. 151122–151152, 2024, doi: 10.1109/ACCESS.2024.3476960.

[8] “Speech emotion recognition using deep learning: From basic to complex emotions in unimodal and multimodal frameworks,” *Neural Computing and Applications*, 2026, doi: 10.1007/s00521-026-12186-w.

[9] S. R. Gudivada, P. Polipogu, R. Kodali, et al., “Speech emotion recognition in adults and children: A comprehensive review of traditional features and raw waveform models,” *International Journal of Speech Technology*, 2025, doi: 10.1007/s10772-025-10229-6.

[10] H. H. Mustafa, N. R. Darwish, and H. A. Hefny, “Automatic speech emotion recognition: A systematic literature review,” *International Journal of Speech Technology*, vol. 27, no. 1, pp. 267–285, 2024, doi: 10.1007/s10772-024-10096-7.

[11] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, “Speech emotion recognition using artificial intelligence: A systematic review and meta-analysis,” *Artificial Intelligence Review*, vol. 58, no. 7, art. 198, 2025, doi: 10.1007/s10462-025-11197-8.

[12] F. S. Alamri, A. R. Mirdad, T. Saba, M. A. Raza, et al., “Impact of combined spectral and prosodic features on speech emotion recognition,” *IEEE Journal*, 2026.

[13] M. Rezaul, F. R. Soha, and T. A. Ananna, “MFCC features for Bangla audio emotion classification,” *AI Journal*, 2026.

[14] S. M. Tasin, M. E. H. Chowdhury, S. Pedersen, et al., “Ensemble machine learning for inner speech recognition,” IEEE Xplore, 2026.

[15] S. Nath, A. K. Shahi, T. Martin, N. Choudhury, and R. Mandal, “Speech emotion recognition using machine learning: A comparative analysis,” *SN Computer Science*, vol. 5, art. 390, 2024, doi: 10.1007/s42979-024-02656-0.

[16] S. Vijayalakshmi and K. R. Kavitha, “Speech emotion recognition using deep learning,” in *Advances in Artificial Intelligence*, book chapter, 2026.

[17] S. G. Prasad and M. Sreevani, “Enhancing speech emotion recognition with deep learning,” *IJSRT Journal*, 2026.

[18] M. Kar, S. Sagnika, and S. Bilgaiyan, “High-accuracy speech emotion recognition using MFCC and optimized LSTM,” Springer book chapter, 2026, doi: 10.1007/978-3-032-14757-8_48.

[19] O. Adelekun et al., “Speech emotion recognition using MFCC and LSTM deep learning model,” arXiv preprint, 2026.

[20] A. I. Middya, B. Nag, and S. Roy, “Effective MLP and CNN based ensemble learning for speech emotion recognition,” *Multimedia Tools and Applications*, 2024.

[21] “Speech emotion recognition using transfer learning and self-supervised speech representation learning,” in Proc. IEEE Conf. [Online]. Available: https://ieeexplore.ieee.org/document/10334799

[22] W. Chen, X. Xing, P. Chen, and X. Xu, “Vesper: A compact and effective pretrained model for speech emotion recognition,” *IEEE Trans. Affective Computing*, vol. 15, no. 3, pp. 1711–1724, 2024, doi: 10.1109/TAFFC.2024.3369726.

[23] S. Y. Chowdhury, B. Banik, M. T. Hoque, and S. Banerjee, “Hybrid deep learning for speech emotion recognition using feature engineering,” arXiv:2507.07046, 2026.

[24] “Efficient speech emotion recognition via spectral learning and attention,” arXiv:2507.03251, 2026.

[25] “Machine learning approach of speech emotions recognition using feature fusion technique,” *Multimedia Tools and Applications*, 2024, doi: 10.1007/s11042-023-16036-y.

[26] “Emotion analysis using deep learning on four datasets,” in Proc. IEEE Conf., 2026.

[27] T. B. Santoso, A. N. Budianto, and T. Dutono, “Building a speech emotion dataset for the Indonesian language,” *IEEE Access*, 2026.

[28] S. Nagro, “Optimization of speech emotion recognition using hybrid dataset integration and deep learning-based feature fusion with a novel balanced focal entropy loss,” *Scientific Reports*, vol. 16, art. 17447, 2026, doi: 10.1038/s41598-026-48975-5.

[29] M. Hossain, M. S. A. Patwary, M. M. Hossain, et al., “Real-time speech emotion recognition with adaptive learning for Bangla,” *SN Computer Science*, 2026.

[30] M. P. Sunil, S. A. Hariprasad, Z. Abdullah, et al., “CNNs and SVM for AI-driven speech emotion recognition,” in Proc. HUMAN 2025, 2026.

[31] G. Muhammad, S. Almuntasheri, F. Alenezi, et al., “EEG-based multimodal emotion recognition: A review,” *ACM Transactions*, 2026.

[32] C. Barhoumi and Y. BenAyed, “Transformer encoder and data augmentation for real-time speech emotion recognition,” *Multimedia Tools and Applications*, 2026, doi: 10.1007/s11042-026-21395-3.

[33] C. Barhoumi and Y. BenAyed, “Real-time speech emotion recognition using deep learning and data augmentation,” *Artificial Intelligence Review*, vol. 58, art. 49, 2025, doi: 10.1007/s10462-024-11065-x.

[34] J. H. Chowdhury, S. Ramanna, and K. Kotecha, “Speech emotion recognition with light weight deep neural ensemble model using hand crafted features,” *Scientific Reports*, vol. 15, art. 11824, 2025, doi: 10.1038/s41598-025-95734-z.