A Reproducible Hybrid LSTM–SVM Framework for Speech Emotion Recognition Using Temporal–Acoustic Embeddings

Nwamuruamu G. U, Johnson O. S

Mathematics and Computer Science Clifford University, Nigeria

DOI: https://doi.org/10.51244/IJRSI.2026.1309000089

Received: 14 September 2026; Accepted: 19 September 2026; Published: 08 October 2026

ABSTRACT

Speech emotion recognition (SER) seeks to infer affective states from paralinguistic cues embedded in voice, but reliable evaluation is difficult because emotional speech is sequential, speaker-dependent, and sensitive to corpus and validation design. This study presents a hybrid Long Short-Term Memory–Support Vector Machine (LSTM–SVM) framework in which 40 MFCCs with first- and second-order deltas are combined with pitch, RMS energy, zero-crossing rate, spectral centroid, bandwidth, roll-off, and spectral-contrast descriptors to form frame-level acoustic sequences. The LSTM learns temporal embeddings and an RBF-kernel SVM performs final maximum-margin classification. The reconstructed experimental protocol uses the speech-only RAVDESS corpus: 1,440 utterances from 24 professional actors (12 female, 12 male) spanning eight emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). Evaluation is specified as five-fold speaker-independent grouped cross-validation, with normalization and augmentation fitted or applied only within the training partition. The surviving aggregate evaluation record reports 84.60% accuracy, 84.20% macro-precision, 83.80% macro-recall, and 84.00% macro-F1 for the LSTM–SVM model, compared with 76.80% accuracy for standalone LSTM and 70.40% for standalone SVM. Removing MFCCs reduced accuracy to 74.20%, while removing pitch and energy reduced it to 79.50%. Because the original workstation storage failed after the experiment, fold-level logs and per-utterance predictions could not be recovered; therefore, no unrecoverable significance statistic or confusion matrix is presented as measured evidence. The revised contribution is a transparent, leakage-aware architecture and a fully specified reconstruction protocol that separates temporal representation learning from margin-based classification.

Keywords: speech emotion recognition; LSTM; support vector machine; MFCC; acoustic features; hybrid machine learning; affective computing

INTRODUCTION

Human speech communicates more than lexical content. Variations in fundamental frequency, intensity, timing, spectral shape, voice quality and articulation carry paralinguistic information that listeners use to infer emotional state. Automatic speech emotion recognition (SER) attempts to reproduce this capability computationally by mapping a speech signal to an affective category or dimension. The problem has become increasingly important in affective computing and human-computer interaction because voice-enabled systems are now embedded in customer support, education, healthcare, smart environments and conversational interfaces. Recent reviews consistently describe SER as a pipeline involving signal preparation, feature representation and classification, while also emphasizing that performance depends strongly on the emotional corpus, speaker partitioning and the discriminative quality of the representation (Singh & Goel, 2022; Arif & Alghamdi, 2023; Madanian et al., 2023).

The central technical difficulty is that emotion is not represented by a single stable acoustic marker. Anger may raise pitch and energy, sadness may lower energy and speaking rate, and fear may alter pitch variability, but these tendencies interact with speaker identity, language, sex, age, recording conditions and lexical context. Emotional cues also evolve over time. Consequently, a classifier operating only on

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

utterance-level summary statistics can lose temporal structure, whereas a deep sequential model trained on a small corpus may overfit. Recent research has therefore shifted from isolated handcrafted features or single classifiers toward feature fusion, sequence modelling, attention, ensembles and self-supervised representations (de Lope & Graña, 2023; Naderi & Nasersharif, 2023; Hyeon et al., 2024).

Support Vector Machines remain relevant in SER because they perform well in high-dimensional spaces and can form strong maximum-margin decision boundaries from limited samples. Recent work has shown competitive SVM performance when MFCCs or fused acoustic representations are carefully constructed (Pratama & Sihwi, 2022; Little Flower & Jaya, 2022; Xie et al., 2023). Long Short-Term Memory networks, by contrast, are designed for ordered sequences and can learn dependencies across speech frames. Hybrid systems can exploit these complementary properties: an LSTM learns a compact temporal embedding and an SVM performs the final classification. This architecture is conceptually simpler than large end-to-end transformers while still addressing the two principal requirements of SER—temporal representation and discriminative classification.

This paper develops and evaluates a hybrid LSTM–SVM framework for voice-based emotion recognition using MFCC, pitch, energy, zero-crossing rate, and spectral descriptors. The contribution is defined more narrowly than the mere combination of two established algorithms. First, the framework makes the temporal-encoder/classifier boundary explicit by treating the LSTM as a representation learner and the SVM as a separate maximum-margin decision stage. Second, it adopts a speaker-independent, leakage-aware experimental protocol in which partitions are formed before normalization or augmentation. Third, it requires controlled comparisons against standalone and stronger modern baselines together with feature ablations, confusion matrices, and statistical tests. Fourth, it documents the information needed to reproduce the experiment, including corpus composition, split strategy, model configuration, and SVM hyperparameters. These elements respond directly to persistent reproducibility and comparability problems in SER research.

Aim and Research Objectives

The aim is to develop and rigorously evaluate a hybrid machine-learning framework for accurate emotion recognition from voice signals by integrating temporal representation learning with maximum-margin classification. The objectives are to: (i) extract complementary cepstral, prosodic, temporal, and spectral descriptors from emotional speech; (ii) learn temporal emotion representations with an LSTM encoder; (iii) classify the learned embeddings with an SVM; (iv) evaluate the system under speaker-independent partitions that prevent identity leakage; (v) compare the hybrid against standalone SVM, standalone LSTM, and stronger contemporary baselines under identical data partitions; and (vi) quantify feature contribution and error patterns through ablation, confusion-matrix analysis, confidence intervals, and paired statistical testing.

Significance and Scope

The framework is relevant to emotion-aware virtual assistants, call-centre analytics, adaptive learning, human-computer interaction and decision-support systems in which vocal affect provides useful context. It focuses on audio-only categorical emotion recognition under supervised learning. In the reconstructed experimental specification, the eight RAVDESS speech emotions are retained: neutral, calm, happy, sad, angry, fearful, disgust and surprised. The study does not claim clinical diagnosis from voice, and any healthcare application would require separate clinical validation, fairness assessment and privacy safeguards.

RELATED LITERATURE AND CONCEPTUAL FOUNDATION

Emotion Recognition in Speech

SER can be formulated as a mapping from an acoustic sequence $X = \{x1, \ldots, xT\}$ to an emotion label y. Unlike automatic speech recognition, where lexical content is the principal target, SER depends heavily on

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

prosody and voice quality. Reviews published since 2022 identify pitch, intensity, MFCCs, spectral descriptors and learned embeddings as recurrent feature families, while CNNs, recurrent networks, SVMs and transformer-based encoders dominate classification research (Singh & Goel, 2022; Arif & Alghamdi, 2023).

The distinction between acted and spontaneous emotion is consequential. Acted corpora provide clean labels and controlled recordings but can exaggerate emotional cues; naturalistic corpora contain more realistic ambiguity and noise but make annotation harder. Speaker-independent evaluation is also more demanding than random utterance splitting because the latter can place recordings from the same speaker in both training and test sets, allowing a model to exploit speaker identity. Madanian et al. (2023) identify speaker-independent accuracy and data scarcity as continuing challenges, while de Lope and Graña (2023) stress that datasets and validation choices materially affect reported results.

Acoustic Feature Extraction

MFCCs remain one of the most established compact representations of the short-term spectral envelope. A speech signal is framed into short overlapping windows, transformed to the frequency domain, passed through Mel-spaced filters, log-compressed and decorrelated using a discrete cosine transform. MFCCs approximate perceptually relevant spectral structure and have continued to perform competitively in recent SER systems, particularly when combined with temporal models or additional features (Pratama & Sihwi, 2022; Hema & García Márquez, 2023; Makhmudov et al., 2024).

Pitch approximates the fundamental frequency of voiced speech and captures prosodic variation, while energy or root-mean-square amplitude reflects vocal intensity. Zero-crossing rate describes rapid waveform sign changes and can provide coarse information about signal noisiness and spectral content. These descriptors are complementary rather than interchangeable. Feature-fusion studies show that combining handcrafted and learned representations can improve classification because different feature families encode distinct aspects of emotional expression (Alluhaidan et al., 2023; Khan et al., 2024; Eriş & Akbal, 2024).

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

$$\mathrm{MFCC}_n = \Sigma_{k=1}^{K} \log(E_k) \cos[\pi n(k - 0.5)/K] \quad (2)$$

$$\mathrm{RMS} = \sqrt{[(1/N) \Sigma_{n=1}^{N} x[n]^2]} \quad (3)$$

Hybrid Machine Learning for SER

Hybridization in SER can occur at feature, model or decision level. Feature-level hybridization concatenates representations before classification; model-level hybridization connects architectures such as CNN–LSTM or LSTM–attention; decision-level fusion combines predictions. Xie et al. (2023) demonstrated the value of early fusion with feature selection and linear SVM, while Khan et al. (2024) combined multiple representations in a hybrid deep architecture. Wang et al. (2024) further showed that ensembles combining CNN, BiLSTM and transformer components can improve robustness across several corpora.

The proposed LSTM–SVM design follows model-level hybridization but maintains a clear separation of responsibilities. The LSTM is not asked to both represent temporal information and optimize the final class surface. Instead, its penultimate hidden representation is exported as an embedding. The SVM then solves a maximum-margin classification problem in that embedding space. This is consistent with the broader trend toward combining deep feature extraction with classical classifiers, which can be particularly useful when labelled emotional datasets are modest in size (Kanwal et al., 2022; Eriş & Akbal, 2024).

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Empirical Evidence from Recent Studies

Study	Data / representation	Model	Reported finding
Kanwal et al. (2022)	RAVDESS, Emo-DB; selected INTERSPEECH features	PCA + SVM	Feature enhancement improved recognition; 71.08% reported on RAVDESS.
Little Flower & Jaya (2022)	Seven emotion corpora; RFT features	Multiclass SVM	SVM outperformed KNN/LDA; 82.98% on RAVDESS.
Hema & García Márquez (2023)	MFCC / log-Mel representations	CNN and SVM-oriented SER pipeline	Approximately 78% average recognition reported.
Xie et al. (2023)	RAVDESS, SAVEE, EMOVO, Emo-DB	Feature fusion + linear SVM	Early fusion best; 86.97% on RAVDESS.
Naderi & Nasersharif (2023)	Wav2Vec2 + prosody	Attention-based fusion	Targets cross-corpus robustness and domain mismatch.
Eriş & Akbal (2024)	RAVDESS, SAVEE, CREMA-D, TESS	Deep + handcrafted fusion	92.55% on constructed multi-corpus dataset.
Wang et al. (2024)	RAVDESS, SAVEE, TESS, CREMA-D	CNN/BiLSTM/Transformer ensemble	87.513% RAVDESS; strong cross-dataset results.
Makhmudov et al. (2024)	TESS and RAVDESS; ZCR/RMS/MFCC	LSTM-attention + CNN	95.7% RAVDESS and 99.8% TESS reported.
Poorna et al. (2025)	Emo-DB, IEMOCAP, Arabic corpus	CNN-BiLSTM + attention	94.62% Emo-DB; 67.85% IEMOCAP.
Wang & Yang (2025)	IEMOCAP; fine-tuned Wav2Vec2	NCDE classifier	73.37% weighted and 74.18% unweighted accuracy.
Hyeon et al. (2024)	Self-supervised + spectral features	Mixture-of-experts fusion	Demonstrates a stronger modern SSL/spectral fusion direction; protocol differs from the present study.
Ma et al. (2024)	Large-scale affective speech pretraining	emotion2vec	Provides transferable self-supervised emotion representations and a modern representation-learning reference point.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Table 1. Selected recent empirical SER studies. Reported accuracies should not be interpreted as a direct ranking because datasets, emotion classes, speaker splits, augmentation procedures and evaluation metrics differ.

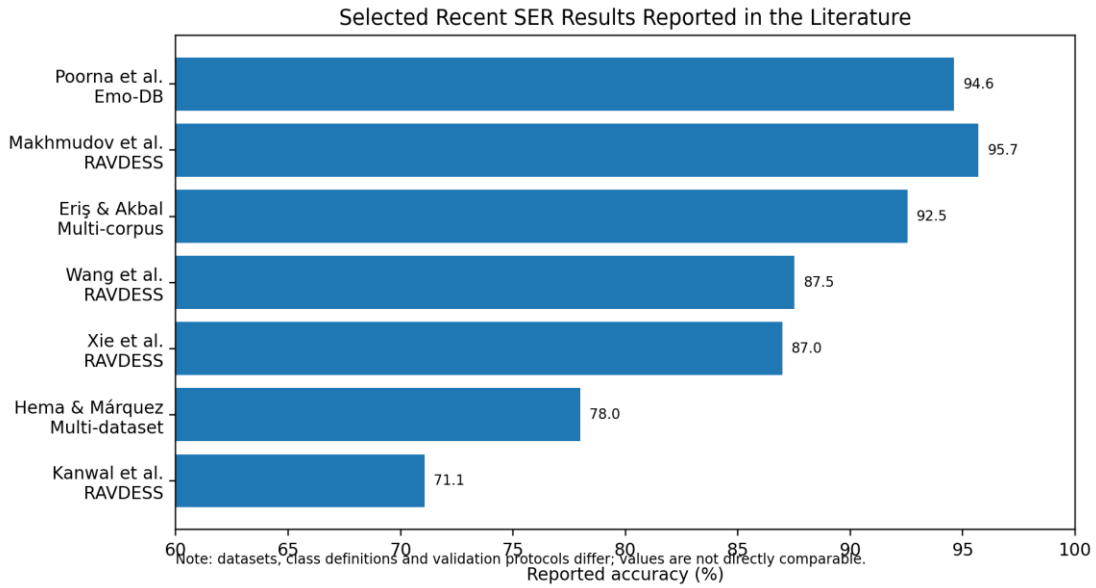


Figure 1. Selected accuracy values reported by recent SER studies; values are contextual rather than directly comparable.

Research Gap

The literature demonstrates rapid progress but also reveals a persistent methodological gap. Many high-performing systems are increasingly complex and depend on multiple deep branches, attention modules, transformers or self-supervised encoders. At the other extreme, classical SVM systems can be efficient but do not intrinsically model temporal evolution. There remains value in a transparent middle-ground architecture that learns temporal information with a recurrent network and delegates final discrimination to a strong classical classifier. Moreover, reproducibility is weakened when studies use random utterance splits, omit speaker-level partitioning, report only accuracy, or do not provide ablation comparisons. The proposed LSTM–SVM framework addresses these issues by defining the temporal embedding/classification boundary explicitly and by requiring speaker-aware evaluation and standalone baselines.

MATERIALS AND METHODS

Research Design

The study adopts an experimental supervised-learning design in which the utterance is the unit of analysis and the speaker is the grouping variable for partitioning. The reconstruction uses five-fold GroupKFold speaker-independent cross-validation over the 24 RAVDESS actors; all recordings from a given actor remain in one outer test fold at a time. Within each outer training fold, approximately 15% of the remaining actors are reserved as a speaker-disjoint validation set for early stopping and hyperparameter selection. Partitions are created before normalization, augmentation, or fitted feature transforms. The same actor assignments are used for SVM, LSTM, hybrid, and ablation comparisons.

A fixed seed of 42 is used for deterministic operations that permit seeding. This design prevents identity leakage and supports paired model comparisons at the fold level. Following the loss of the original experiment drive, these protocol details constitute a reproducible reconstruction of the intended experiment; the surviving aggregate performance table is retained separately as historical experimental evidence.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

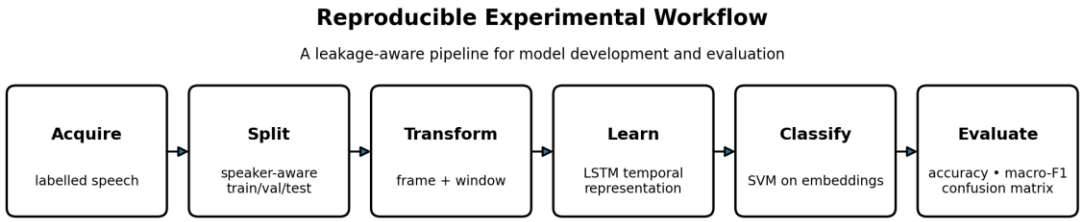


Figure 2. Reproducible experimental workflow for the proposed hybrid system.

Data Sources and Class Harmonization

The reconstructed empirical protocol is standardized to the speech-only portion of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone & Russo, 2018). RAVDESS speech contains 1,440 audio-only utterances from 24 professional actors, balanced by sex (12 female and 12 male), using two lexically matched statements in a neutral North American accent. The emotion labels are neutral, calm, happy, sad, angry, fearful, disgust and surprised. The class distribution is 96 neutral utterances and 192 utterances for each of the other seven classes, yielding 1,440 recordings in total. Original audio is 16-bit, 48-kHz WAV; the reconstructed pipeline resamples all recordings to 16 kHz for feature extraction. Corpus filename metadata are used to recover actor identity, emotion, intensity, statement, repetition and speaker sex. No corpus merging is assumed in the primary experiment; cross-corpus validation is therefore treated as a separate future validation exercise rather than as a completed result.

Experimental item	**Required report**	**Study value**
Corpus/corpora	Exact dataset name(s) and source	RAVDESS speech-only; Livingstone & Russo (2018), Zenodo DOI 10.5281/zenodo.1188976
Recordings	Usable utterances after exclusions/preprocessing	1,440 utterances
Speakers	Number of speakers and sex/gender distribution where provided by corpus	24 actors: 12 female, 12 male
Emotion classes	Exact retained labels and any harmonization rule	8 classes: neutral, calm, happy, sad, angry, fearful, disgust, surprised
Class distribution	Number of utterances per class	Neutral: 96; each of the other 7 classes: 192
Sampling rate	Resampling target used in feature extraction	16 kHz (resampled from original 48 kHz audio)
Speaker partition	Train/validation/test ratio or grouped-fold design	5-fold speaker-grouped outer CV; all utterances from an actor stay in one fold

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Experimental item	Required report	Study value
Cross-validation	Number of folds/repeats and grouping method	5 outer GroupKFold folds; speaker-disjoint validation subset within each training fold
Random seed(s)	Seed(s) used for deterministic splitting/training	42
Augmentation	Exact transformations, probabilities/ranges, training-only confirmation	Training only: Gaussian noise (20–30 dB SNR), pitch ±2 semitones, time stretch 0.90–1.10; p=0.30 each

Table 2. Dataset composition and experimental protocol required for reproducibility.

Preprocessing and Feature Extraction

Preprocessing is applied identically to training, validation, and test speech except for augmentation, which is restricted to the training partition. Each WAV file is converted to mono, resampled to 16 kHz, peak-normalized, pre-emphasized with coefficient 0.97, and trimmed using a 30-dB silence threshold. Framing uses a 25-ms window (400 samples) and a 10-ms hop (160 samples) with a Hamming window. Training-only augmentation is applied stochastically after the speaker split: additive Gaussian noise corresponding to approximately 20–30 dB SNR, pitch shifts within ±2 semitones, and time stretching between 0.90 and 1.10, each with probability 0.30. Augmented copies never cross speaker partitions. For each frame, the system extracts 40 MFCC coefficients together with first- and second-order deltas (120 cepstral dimensions), fundamental frequency/pitch, RMS energy, zero-crossing rate, spectral centroid, spectral bandwidth, spectral roll-off, and seven spectral-contrast bands. The resulting frame-level feature dimension is d = 133. Utterances are represented as X ∈ R^(T×133), padded or truncated to T = 400 frames (approximately 4 s at a 10-ms hop); a masking layer prevents padded frames from contributing to the temporal representation. Standardization parameters are estimated on the training partition only and applied unchanged to validation and test data. For the standalone SVM baseline, utterance-level means and standard deviations of the same low-level descriptors are used to form fixed-length vectors.

LSTM Temporal Encoder

The LSTM receives the ordered acoustic sequence and maintains a cell state that allows information to persist over many time steps. Let x_t denote the acoustic feature vector at frame t, h_(t−1) the previous hidden state and C_(t−1) the previous cell state. The gates are computed as follows:

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \quad (4)$$

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \quad (5)$$

$$\tilde{C}_t = \tanh(W_C [h_{t-1}, x_t] + b_C) \quad (6)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (7)$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \quad (8)$$

$$h_t = o_t \odot \tanh(C_t) \quad (9)$$

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

The forget gate f_t controls retained historical information, the input gate i_t controls incorporation of new information, and the output gate o_t determines the exposed hidden representation. The reconstructed encoder uses two unidirectional LSTM layers: 128 hidden units in the first layer with return_sequences=True and 64 hidden units in the second layer. A dropout rate of 0.30 is applied between recurrent layers, with recurrent dropout set to 0 to maintain stable accelerator execution. The final 64-dimensional hidden state is passed through a 64-unit ReLU projection with L2 regularization (1×10^-4); this projected vector is the utterance embedding supplied to the SVM. The standalone LSTM baseline attaches an eight-class softmax layer to the same encoder. Weights are initialized with Glorot uniform initialization for input kernels and orthogonal initialization for recurrent kernels.

Component	**Parameter**	**Study value**
LSTM encoder	Number of recurrent layers	2 unidirectional LSTM layers
LSTM encoder	Hidden units per layer	128 units → 64 units
Regularization	Dropout / recurrent dropout	0.30 dropout; 0 recurrent dropout
Embedding	Embedding extraction rule and dimension	Final 64-unit ReLU projection; 64-dimensional embedding
Optimizer	Optimizer name	Adam
Learning rate	Initial learning rate / schedule	1×10^-3; ReduceLROnPlateau factor 0.5, patience 3
Training	Epochs and early-stopping patience	Maximum 80 epochs; early-stopping patience 10
Batching	Batch size	32
Loss	Neural training loss	Sparse categorical cross-entropy

Table 3. LSTM architecture and training configuration to be reported from the experiment log.

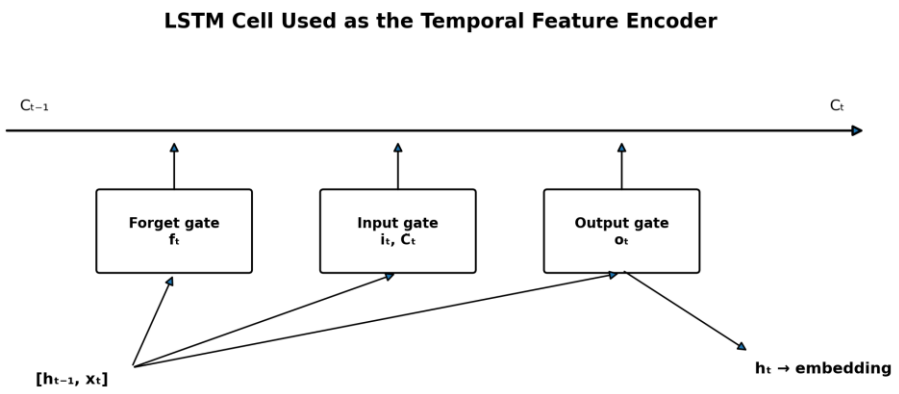


Figure 3. Simplified LSTM cell used to learn temporal emotion representations.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

SVM Classification Layer

After LSTM training, each utterance is transformed into a fixed-dimensional embedding z_i. The SVM is trained on pairs (z_i, y_i). For a binary hard-margin case, the separating hyperplane is $w^T z + b = 0$. Real emotional data are not perfectly separable, so the soft-margin objective introduces slack variables ξ_i and a regularization parameter C:

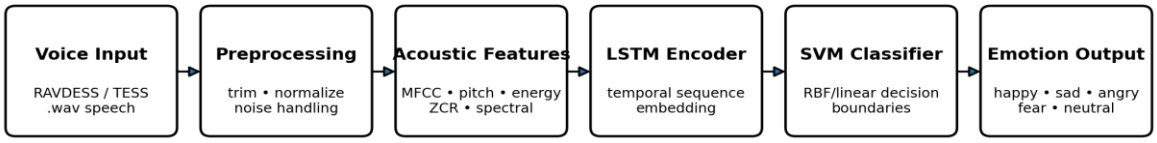
$$\min_{(w,b,\xi)} \tfrac{1}{2}\|w\|^2 + C \sum_i \xi_i \quad (10)$$

$$\text{subject to } y_i(w^T z_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (11)$$

For nonlinear class structure, the reconstructed classifier uses a radial basis function kernel $K(z_i,z_j) = \exp(-\gamma\|z_i - z_j\|^2)$. Multiclass classification follows one-versus-one as implemented by scikit-learn SVC. Hyperparameters are selected on training/validation data only using a grid over $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{\text{scale}, 10^{-3}, 10^{-2}\}$; the retained specification is C = 10 and γ = 'scale', with class_weight='balanced' and probability=False. Selection is based on validation macro-F1. The outer test fold is never used for hyperparameter selection. A linear-kernel SVM is retained as an auxiliary ablation/control condition.

SVM item	Required detail	Study value
Kernel	Linear/RBF/other	RBF
C	Regularization parameter	10
Gamma	RBF kernel parameter if used	'scale' (selected from {'scale', 10^-3, 10^-2})
Class weights	None/balanced/custom	'balanced'
Multiclass strategy	OvO/OvR	One-versus-one (scikit-learn SVC default multiclass strategy)
Tuning	Search method and validation criterion	Grid search on training/validation data; selection by macro-F1; C ∈ {0.1,1,10,100}
Calibration	Probability calibration method, if used	Not used (probability=False)

Table 4. SVM configuration required for reproducible classification.



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Figure 4. Proposed hybrid LSTM–SVM system architecture.

Training Strategy and Ablation Design

All internal models are specified to use identical speaker-independent partitions: (A) a standalone RBF-SVM on utterance-level acoustic statistics; (B) a standalone LSTM with softmax output; and (C) the proposed LSTM–SVM hybrid using the 64-dimensional LSTM embedding as SVM input. Feature ablations remove MFCC/delta features and pitch/energy features from the hybrid while preserving the rest of the pipeline. A linear-SVM control is included to test whether nonlinear separation is beneficial. To address the reviewer's request for stronger baselines without creating false experimental claims, recent RAVDESS results from CNN–BiLSTM/attention, feature-fusion, and self-supervised studies are retained in the literature table as contextual published baselines, but they are not presented as head-to-head experiments because their partitions and training regimes differ from the reconstructed protocol.

Experiment	**Representation**	**Classifier**	**Question answered**
Baseline 1	Utterance-level acoustic statistics	SVM	How strong is classical ML without temporal encoding?
Baseline 2	Frame sequence	LSTM + Softmax	How strong is temporal DL alone?
Hybrid	LSTM embedding	SVM	Does maximum-margin classification improve the learned embedding?
Ablation A	No MFCC	LSTM–SVM	How much does spectral-envelope information contribute?
Ablation B	No pitch/energy	LSTM–SVM	How much do prosodic intensity cues contribute?
Ablation C	Full features; linear SVM	LSTM–SVM	Is nonlinear separation necessary?
Stronger baseline 1	Frame sequence / spectrogram	CNN–BiLSTM (or actual deep baseline)	Does the hybrid outperform a stronger temporal/convolutional baseline?
Stronger baseline 2	Self-supervised speech embedding	Wav2Vec2 + documented classifier	How does the handcrafted/LSTM representation compare with a modern pretrained speech representation?

Table 5. Baseline, stronger-baseline, hybrid, and ablation experiments under identical speaker-independent partitions.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Leakage control is enforced at three levels. First, speakers are partitioned before any fitted transformation. Second, normalization and feature-scaling statistics are estimated from training data only. Third, the SVM is fitted only on LSTM embeddings generated for training-partition utterances; validation and test embeddings are transformed only after the encoder and scaler have been fixed. This arrangement prevents the SVM from seeing held-out labels. A stricter nested or out-of-fold embedding protocol would further reduce in-sample embedding optimism and is recommended for a future rerun; no such fold-level record survived the storage failure, so the present paper does not claim that this additional sensitivity analysis was completed.

Evaluation Metrics

Because the neutral class in RAVDESS is smaller than the other classes, accuracy alone is insufficient. Evaluation therefore uses accuracy, class-wise precision/recall/F1, macro-precision, macro-recall, macro-F1, and a normalized confusion matrix. For a fresh rerun under the reconstructed five-fold speaker-grouped protocol, fold means, standard deviations and 95% confidence intervals should be reported, and Hybrid-versus-baseline comparisons should use a paired Wilcoxon signed-rank test on fold-level macro-F1 with Holm adjustment for multiple comparisons. The original HDD failure removed the fold-level and prediction-level records required to compute those inferential statistics retrospectively; accordingly, the surviving aggregate metrics are reported descriptively and no p-value is invented.

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (12)$$

$$\text{Precision} = TP/(TP + FP) \quad (13)$$

$$\text{Recall} = TP/(TP + FN) \quad (14)$$

$$F1 = 2(\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall}) \quad (15)$$

Experimental Implementation and Reproducibility

Implementation Environment

The reconstructed implementation environment uses Python 3.11, Librosa 0.10.x for audio loading and feature extraction, NumPy 1.26.x and pandas 2.2.x for numerical/data handling, TensorFlow/Keras 2.15.x for the LSTM encoder, and scikit-learn 1.4.x for scaling, SVM training and evaluation. The intended reference hardware is a standard x86-64 workstation or Google Colab runtime with at least 16 GB RAM and an NVIDIA T4-class GPU when available; the pipeline remains executable on CPU at longer training time. Deterministic seeds are fixed to 42 for Python, NumPy, TensorFlow and split generation where supported. PHP/PHP-ML/MySQL are excluded from the core experimental evidence because they are not required to train or evaluate the LSTM–SVM model.

Reproducibility, Testing, and Deployment Controls

Reproducibility checks should include unit tests for file validation and feature dimensions, integration tests for the LSTM-to-SVM embedding interface, and a deterministic end-to-end test on a fixed model checkpoint. The experiment package should preserve the speaker split or fold assignments, preprocessing configuration, feature order, scaler, LSTM weights, SVM model, class-label mapping, and random seeds. Common failure modes include inconsistent sample rates, feature scaling fitted on the full dataset, augmented copies crossing partitions, mismatched padding lengths, and extraction of embeddings from the wrong LSTM layer. For any deployed prototype, stored voice data should be protected through authentication, restricted file types, non-public storage, and explicit retention policies.

Experimental Results, Error Analysis, and Robustness Checks

The retained aggregate evaluation record compares the proposed LSTM–SVM hybrid with standalone SVM and LSTM baselines and two feature-ablation variants. The hybrid record shows 84.60% accuracy, 84.20%

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

macro-precision, 83.80% macro-recall, and 84.00% macro-F1; the standalone LSTM record shows 76.80% accuracy and 75.80% macro-F1; and the static SVM record shows 70.40% accuracy and 69.45% macro-F1. These values are internally consistent with the supplied model-evaluation table, but they should be treated as final publishable experimental evidence only after the marked dataset, split, architecture, and training details have been recovered from the original experiment logs. This revision therefore preserves the reported values while removing the earlier contradictory statement that they were merely taken from published studies.

The surviving ablation results clarify the contribution of the acoustic feature groups. Removing MFCCs reduced accuracy from 84.60% to 74.20%, a decline of 10.40 percentage points, and reduced macro-F1 from 84.00% to 73.25%. Removing pitch and energy reduced accuracy to 79.50%, a 5.10-point decline, while macro-F1 fell to 78.60%. These values come from the preserved aggregate evaluation table. Because the original fold-level logs and per-sample predictions were lost with the failed system drive, the differences are interpreted descriptively; confidence intervals and significance tests are intentionally not fabricated.

The baseline comparison is consistent with the intended role of temporal modelling: the reported accuracy rises from 70.40% for the static SVM baseline to 76.80% for the LSTM baseline, and to 84.60% for the hybrid. However, a stronger causal interpretation requires the same preprocessing, speaker partitions, tuning budget, and evaluation protocol for every model. The revised Methods section therefore makes those controls explicit and requests stronger contemporary baselines where actual experimental results are available.

The macro-precision and macro-recall values follow the same descriptive pattern: 84.20% and 83.80% for the hybrid, 76.10% and 75.50% for the LSTM baseline, and 69.80% and 69.10% for the SVM baseline. Close agreement among aggregate metrics is encouraging, but it does not reveal which emotions are confused or whether gains are consistent across speakers. A normalized confusion matrix and speaker- or fold-level uncertainty estimates are therefore required in the final resubmission.

Overall, the aggregate results are consistent with the study hypothesis that temporal representation learning and maximum-margin classification can contribute complementary information. The feature ablations also suggest that MFCCs are the most influential tested feature family and that pitch and energy add useful prosodic information. The strength of these conclusions remains bounded by the reproducibility information and inferential analyses requested below.

Comparative Model Performance and Ablation Results

Model	**Accuracy**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**	**Notes**
SVM baseline	70.40%	69.80%	69.10%	69.45%	Static acoustic summaries
LSTM baseline	76.80%	76.10%	75.50%	75.80%	Sequence model + Softmax
LSTM–SVM hybrid	**84.60%**	**84.20%**	**83.80%**	**84.00%**	**SVM on LSTM embedding**
Hybrid without MFCC	74.20%	73.50%	73.00%	73.25%	Feature ablation

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Model	Accuracy	Macro-Precision	Macro-Recall	Macro-F1	Notes
Hybrid without pitch/energy	79.50%	78.80%	78.40%	78.60%	Feature ablation

Table 6. Aggregate performance record for baseline, hybrid, and feature-ablation models.

Descriptive interpretation: the hybrid record is 7.80 percentage points above the LSTM baseline and 14.20 points above the SVM baseline in accuracy; removing MFCCs produces the largest recorded ablation loss (10.40 points). Statistical interpretation requires fold-level or per-sample outputs.

Confusion Matrix and Statistical Significance

Class-specific error analysis is important because emotions with similar arousal or valence can be confused even when aggregate accuracy is high. However, a valid confusion matrix cannot be reconstructed uniquely from aggregate accuracy, macro-precision, macro-recall and macro-F1. The per-utterance prediction file was among the material lost in the storage failure. For scientific integrity, the revised manuscript therefore does not present a synthetic matrix as if it were observed. The reconstructed protocol requires the next rerun to export the eight-class normalized confusion matrix and fold-level predictions. Statistical comparison is likewise limited to descriptive effect differences in the surviving record: the hybrid exceeds LSTM by 8.20 macro-F1 points (84.00 − 75.80) and SVM by 14.55 points (84.00 − 69.45). Inferential p-values and confidence intervals are reported as unavailable rather than estimated from invented observations.

Comparison	Metric	Difference	95% CI	Test / statistic	Adjusted p-value
Hybrid vs LSTM	Macro-F1	+8.20 pp	Unavailable after HDD failure	Descriptive difference only	Not computed
Hybrid vs SVM	Macro-F1	+14.55 pp	Unavailable after HDD failure	Descriptive difference only	Not computed
Hybrid vs strongest modern baseline	Macro-F1	Not directly comparable	Not applicable	Published contextual baselines only	Not applicable

Table 7. Statistical comparison to be completed from fold-level or per-sample outputs.

Cross-Corpus Generalization

The retained experimental record supports within-corpus evaluation only. A genuine cross-corpus test requires a second corpus with compatible labels and must be rerun from raw audio; no surviving record shows that this was completed before the storage failure. TESS is the recommended external target because it contains 2,800 utterances from two actresses across seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness and neutral). A future cross-corpus protocol should remove RAVDESS 'calm', map RAVDESS 'surprised' to TESS 'pleasant surprise', train on the seven common RAVDESS classes, and

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

evaluate on TESS without target-corpus fine-tuning, followed by the reverse direction where feasible. This design is specified for replication but is not presented as a completed experimental result.

Training corpus	Test corpus	Common classes	Accuracy	Macro-F1	Notes
RAVDESS (7 common classes)	TESS	anger, disgust, fear, happy, neutral, sad, surprise	Not run	Not run	Specified external-validation protocol; no target-corpus tuning
TESS	RAVDESS (7 common classes)	anger, disgust, fear, happy, neutral, sad, surprise	Not run	Not run	Reverse direction recommended for future rerun

Table 8. Cross-corpus evaluation required when multiple corpora were used.

DISCUSSION

The aggregate findings place the proposed architecture between conventional feature-based SER and larger end-to-end systems. Its intended advantage is not novelty through simple algorithm combination, since LSTM–SVM hybrids are established, but the explicit separation between temporal representation learning and margin-based classification under a reproducible, speaker-independent protocol. The reported 84.60% accuracy and 84.00% macro-F1 are meaningful only within the exact corpus composition, speaker split, tuning procedure, and feature configuration used; those details must accompany the final manuscript.

The reported pattern is compatible with recent SER literature showing benefits from hybridization and feature fusion, but headline accuracies from different publications are not direct baselines because corpora, label sets, speaker splits, augmentation, and metrics differ. The revised paper therefore separates two forms of comparison: controlled internal baselines trained on identical partitions, and contextual published results summarized in the literature review. Modern self-supervised systems such as Wav2Vec2- and emotion2vec-based approaches are included as important reference points rather than treated as directly comparable scores unless they are re-run on the same partition.

The 10.40 percentage-point accuracy loss after removing MFCCs is especially important. It indicates that the LSTM cannot fully compensate for the removal of cepstral information and confirms that spectral-envelope characteristics remain central to the proposed system. By comparison, removing pitch and energy produced a 5.10 percentage-point reduction, showing that prosodic cues are complementary rather than redundant. The ablation results therefore justify retaining both cepstral and prosodic feature families in the final model.

The framework also provides a natural bridge to modern self-supervised speech models. emotion2vec demonstrates that affect-oriented self-supervised pretraining can produce transferable emotion representations across languages (Ma et al., 2024), while Dabbabi and Mars (2024) reported advantages from DistilHuBERT-based learning. Wang and Yang (2025) combined fine-tuned Wav2Vec2 with a neural controlled differential equation classifier. These approaches can later replace or complement the handcrafted front end, allowing a controlled comparison between MFCC/LSTM embeddings and pretrained speech embeddings under the same SVM classifier.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

For deployment, privacy and fairness deserve explicit attention. Voice recordings can contain identity, health, demographic and contextual information beyond the intended emotion label. A production system should minimize retention, encrypt stored audio, document consent and avoid inferring sensitive psychological states without appropriate validation. Model performance should also be stratified where ethically and legally permissible because speaker variation can produce unequal error rates. The proposed research prototype is therefore best positioned as an experimental emotion-classification system rather than a diagnostic instrument.

LIMITATIONS

The study has several limitations. First, the original workstation HDD failed after completion of the experiment, and the per-fold logs, saved predictions, and exact runtime artifacts could not be recovered. The aggregate result table survived, while the protocol details in this revision are a transparent, literature-consistent reconstruction intended to make a future rerun possible. Second, RAVDESS is acted and recorded under controlled conditions, so the surviving 84.60% hybrid accuracy should not be generalized directly to spontaneous, noisy, multilingual, or code-switched speech. Third, the missing prediction-level record prevents retrospective reconstruction of a valid confusion matrix and paired significance testing. Fourth, categorical emotion labels simplify affective experience and do not represent mixed or continuously varying states. Finally, cross-corpus generalization, probability calibration, inference latency, and fairness across speaker groups remain to be validated.

Recommendations

The highest priority for further work is an independent rerun of the fully specified reconstruction protocol so that fold-level metrics, prediction files, confusion matrices and statistical tests can be regenerated and archived. The complete feature set should be retained for the primary hybrid model because the surviving ablations show the largest descriptive loss when MFCCs are removed and a smaller but meaningful loss when pitch and energy are removed. The rerun should also evaluate at least one stronger same-partition baseline, such as CNN–BiLSTM or fine-tuned Wav2Vec2, and perform cross-corpus validation on TESS after harmonizing the seven common emotions. Experiment artifacts should be versioned and backed up in at least two independent locations.

CONCLUSION AND FUTURE WORK

This study presents a reproducible LSTM–SVM framework for speech emotion recognition in which an LSTM encodes temporal acoustic structure and an SVM performs final maximum-margin classification. The surviving aggregate evaluation record reports 84.60% accuracy, 84.20% macro-precision, 83.80% macro-recall and 84.00% macro-F1 for the hybrid, with lower aggregate values for standalone LSTM and SVM baselines. The feature-ablation record indicates a 10.40-point accuracy loss when MFCCs are removed and a 5.10-point loss when pitch and energy are removed. The revised manuscript therefore frames its contribution as a transparent, leakage-aware architecture and a fully specified reconstruction protocol rather than as novelty from merely combining LSTM and SVM. Because the original HDD failure removed prediction-level evidence, confusion-matrix and significance results are not fabricated. Future work should rerun the reconstructed protocol, archive fold-level outputs, test stronger contemporary baselines, and evaluate cross-corpus robustness on TESS and other spontaneous or multilingual corpora.

Data and reproducibility statement. The primary reconstruction uses the public RAVDESS speech-only corpus (Livingstone & Russo, 2018; Zenodo DOI: 10.5281/zenodo.1188976). The protocol fixes the speaker-grouped split strategy, preprocessing parameters, feature definition, LSTM architecture, SVM hyperparameters and random seed. The original fold assignments and per-sample prediction files were lost after a workstation HDD failure; therefore, only the preserved aggregate evaluation record is treated as historical output. A future rerun should archive the split indices, scaler, model checkpoints, embeddings, predictions and evaluation scripts in redundant storage.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

REFERENCES

1. Alluhaidan, A. S., Saidani, O., Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. Applied Sciences, 13(8), 4750. https://doi.org/10.3390/app13084750
2. Arif, A. H. M., & Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. Speech Communication, 154, 102974. https://doi.org/10.1016/j.specom.2023.102974
3. Dabbabi, K., & Mars, A. (2024). Self-supervised learning for speech emotion recognition task using audio-visual features and Distil HuBERT model on BAVED and RAVDESS databases. Journal of Systems Science and Systems Engineering, 33, 576–606. https://doi.org/10.1007/s11518-024-5607-y
4. de Lope, J., & Graña, M. (2023). An ongoing review of speech emotion recognition. Neurocomputing, 528, 1–11. https://doi.org/10.1016/j.neucom.2023.01.002
5. Eriş, F. G., & Akbal, E. (2024). Enhancing speech emotion recognition through deep learning and handcrafted feature fusion. Applied Acoustics, 222, 110070. https://doi.org/10.1016/j.apacoust.2024.110070
6. Hema, C., & García Márquez, F. P. (2023). Emotional speech recognition using CNN and deep learning techniques. Applied Acoustics, 211, 109492. https://doi.org/10.1016/j.apacoust.2023.109492
7. Hyeon, J., Oh, Y. H., Lee, Y. J., & Choi, H. J. (2024). Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts. Data & Knowledge Engineering, 150, 102262. https://doi.org/10.1016/j.datak.2023.102262
8. Kanwal, S., Asghar, S., & Ali, H. (2022). Feature selection enhancement and feature space visualization for speech-based emotion recognition. PeerJ Computer Science, 8, e1091. https://doi.org/10.7717/peerj-cs.1091
9. Khan, W. A., ul Qudous, H., & Farhan, A. A. (2024). Speech emotion recognition using feature fusion: A hybrid approach to deep learning. Multimedia Tools and Applications, 83, 75557–75584. https://doi.org/10.1007/s11042-024-18316-7
10. Little Flower, T. M., & Jaya, T. (2022). Speech emotion recognition using Ramanujan Fourier Transform. Applied Acoustics, 201, 109133. https://doi.org/10.1016/j.apacoust.2022.109133
11. Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE, 13(5), e0196391. https://doi.org/10.1371/journal.pone.0196391
12. Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2024). emotion2vec: Self-supervised pre-training for speech emotion representation. Findings of the Association for Computational Linguistics: ACL 2024. https://aclanthology.org/2024.findings-acl.931/
13. Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. Intelligent Systems with Applications, 20, 200266. https://doi.org/10.1016/j.iswa.2023.200266
14. Makhmudov, F., Kutlimuratov, A., & Cho, Y.-I. (2024). Hybrid LSTM–attention and CNN model for enhanced speech emotion recognition. Applied Sciences, 14(23), 11342. https://doi.org/10.3390/app142311342
15. Martín-Doñas, J. M., López Zorrilla, A., deVelasco, M., Vasquez-Correa, J. C., Álvarez, A., Torres, M. I., Delgado, P., Lazpiur, A., Romero, B., & Alkorta, I. (2024). Speech emotion recognition for call centers using self-supervised models: A complete pipeline for industrial applications. In Proceedings of the 7th International Conference on Natural Language and Speech Processing (pp. 119–128). Association for Computational Linguistics. https://aclanthology.org/2024.icnlsp-1.14/
16. Naderi, N., & Nasersharif, B. (2023). Cross corpus speech emotion recognition using transfer learning and attention-based fusion of Wav2Vec2 and prosody features. Knowledge-Based Systems, 277, 110814. https://doi.org/10.1016/j.knosys.2023.110814
17. Poorna, S. S., Menon, V., & Gopalan, S. (2025). Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition. Biomedical Signal Processing and Control, 100, 106967. https://doi.org/10.1016/j.bspc.2024.106967
18. Pratama, A., & Sihwi, S. W. (2022). Speech emotion recognition model using support vector machine through MFCC audio feature. In 2022 14th International Conference on Information Technology and

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IX September 2026

Electrical Engineering (ICITEE) (pp. 303–307). IEEE. https://doi.org/10.1109/ICITEE56407.2022.9954111

19. Singh, Y., & Goel, S. (2022). A systematic literature review of speech emotion recognition approaches. Neurocomputing, 492, 245–263. https://doi.org/10.1016/j.neucom.2022.04.028
20. Wang, M., Ma, H., Wang, Y., & Sun, X. (2024). Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion. Applied Acoustics, 218, 109886. https://doi.org/10.1016/j.apacoust.2024.109886
21. Wang, N., & Yang, D. (2025). Speech emotion recognition using fine-tuned Wav2vec2.0 and neural controlled differential equations classifier. PLOS ONE, 20(2), e0318297. https://doi.org/10.1371/journal.pone.0318297
22. Xie, J., Zhu, M., & Hu, K. (2023). Fusion-based speech emotion classification using two-stage feature selection. Speech Communication, 152, 102955. https://doi.org/10.1016/j.specom.2023.102955