Journal of Computational Systems and Applications

https://jcsa.cultechpub.com/jcsa

E-ISSN: 2982-3501

Vol. 3, No. 2, November 2026

DOI: 10.64229/xd5zwc31

Article

Speech Emotion Recognition Using Verbal Behaviour Analysis and a Performance-Weighted Ensemble of Classical Machine Learning Classifiers

Syeda Samra Amir[1,*], Syeda Rabiya Amir[2], Tayyab Ahmed Sheikh[1], Hira Fatima[1], Areshba Mushtaq[1]

[1]Department of Electrical and Computer Systems Engineering, Hamdard University, Karachi, Pakistan

[2]Department of Electronics Engineering, Dawood University of Engineering and Technology, Karachi, Pakistan

[*]Corresponding author: Syeda Samra Amir, syeda.samra@hamdard.edu.pk

Abstract

Objectives/Scope: Speech Emotion Recognition (SER) enables automated systems to infer affective state from vocal cues, with applications in mental health monitoring and human-computer interaction. This study investigates a lightweight, interpretable alternative to costly deep learning, built on classical machine learning and a broadened, Verbal Behaviour Analysis (VBA)-based feature set. Methods, Procedures, and Process: A merged, augmented corpus was constructed from the Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), spanning eight emotion classes. Verbal Behaviour Analysis features Mel-Frequency Cepstral Coefficients (MFCC), pitch, loudness, pause structure, jitter, shimmer, spectral centroid, and zero-crossing rate were extracted, standardized, and reduced via Principal Component Analysis (PCA). Five classifiers (Support Vector Machine [SVM], k-Nearest Neighbors [KNN], Random Forest, XGBoost, and HistGradientBoosting [HGB]) were trained on a recording-grouped, stratified split keeping all augmented variants of a recording together, and combined through a performance-weighted voting ensemble. Results, Observations, and Conclusions: SVM was the strongest individual classifier (93.81% accuracy), followed by KNN (91.98%), HGB (91.36%), Random Forest (88.94%), and XGBoost (88.06%). The weighted ensemble achieved the highest overall accuracy (93.96%) and was most balanced on the most class-imbalanced category, calm. Fifty principal components retained 99.74% of the variance. Novel/Additive Information: This study contributes a Verbal Behaviour Analysis feature set that adds voice-quality descriptors (jitter, shimmer) to spectral and prosodic features without a statistically significant accuracy cost, a recording-grouped validation protocol that avoids data leakage in augmented-then-split corpora, and a transparent, fixed-weight ensemble whose class-level contribution is explicitly quantified.

Keywords

Speech emotion recognition, Verbal behaviour analysis, Machine learning, Weighted ensemble, Prosodic features, Support vector machine

Article History

Received: 11 August 2026 Revised: 19 October 2026

Accepted: 23 November 2026 Available Online: 27 November 2026

Copyright

© 2026 by the authors. This article is published by the Cultech Publishing Sdn. Bhd. under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0): https://creativecommons.org/licenses/by/4.0

1. Introduction

Speech is one of the richest channels of human affective expression, encoding emotional information not only through linguistic content but through prosody, voice quality, and articulatory dynamics. Speech Emotion Recognition (SER) is the computational task of automatically identifying a speaker's emotional state from such vocal cues, and it underlies a growing range of applications, including affective human-computer interaction, call-centre quality monitoring, in-vehicle driver-state assessment, and remote mental health screening. Unlike text-based sentiment analysis, SER must contend with substantial inter-speaker variability, cultural differences in emotional expression, and the difficulty of isolating emotion-relevant acoustic cues from confounding factors such as recording conditions and speaker identity.

Two broad methodological traditions have emerged to address this task. The first relies on handcrafted acoustic and prosodic features, drawn from decades of research in phonetics and voice science, paired with classical machine learning classifiers such as Support Vector Machines and ensemble trees [1,2]. The second, more recent tradition applies deep neural architectures directly to spectrograms or raw waveforms [3,4], allowing the model to learn its own feature representations at the cost of substantially higher computational demand and reduced interpretability. Both traditions have produced strong benchmark results, but the deep-learning tradition in particular has raised a recurring practical concern: architectures capable of state-of-the-art accuracy are often too computationally demanding for real-time or on-device deployment [5,6], which is precisely the setting in which many of SER's most valuable applications mobile mental health monitoring, embedded human-computer interfaces, in-cabin driver monitoring must operate.

Verbal Behaviour Analysis (VBA), as adopted in this study, situates itself within the classical machine learning tradition, but broadens the conventional feature set beyond spectral descriptors alone to jointly capture prosodic cues (pitch, loudness, pause structure) and voice-quality cues (jitter, shimmer) that are established in clinical voice science as correlates of affective and psychological state, and that recent SER work has begun to incorporate directly for this reason [7,8]. The premise of this study is that a sufficiently comprehensive handcrafted feature set can substantially narrow the gap between classical machine learning and more complex ensemble or deep learning approaches, without incurring their computational cost.

Despite considerable progress in this area, a review of the recent literature reveals three persistent gaps, discussed in detail in Section 2: feature sets in most studies remain restricted to spectral and basic prosodic descriptors, with voice-quality measures rarely incorporated alongside them [1,2,9]; ensembling strategies tend either to be absent or to rely on complex, difficult-to-interpret combination rules such as learned gating networks or softmax-learned weighting [10,11]; and reported accuracy figures vary considerably across studies in ways that are attributable to dataset and task differences rather than to genuine methodological superiority, complicating fair comparison, as illustrated by identical-pipeline accuracy swings of over 20 percentage points across corpora [12]. Motivated by these gaps, the present study designs and evaluates a Verbal Behaviour Analysis feature-extraction pipeline, trains and benchmarks five classical machine learning classifiers on a merged, augmented, eight-class emotional speech corpus, combines these classifiers through a transparent, performance-weighted voting ensemble, and benchmarks the resulting approach against recent Speech Emotion Recognition literature using a consistent set of comparison parameters.

The remainder of this paper is organized as follows. Section 2 reviews the related literature and summarizes it in a literature review matrix. Section 3 describes the dataset, feature extraction, and classification methodology. Section 4 presents the experimental results. Section 5 discusses their implications, Section 6 states the limitations of the study, and Section 7 concludes the paper. The overall pipeline, from raw speech corpora to final ensemble evaluation, is summarized in Figure 1.

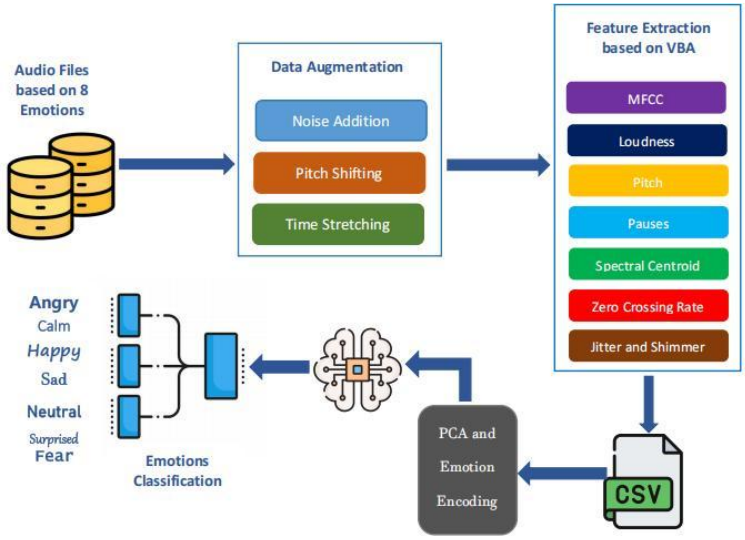


Figure 1. End-to-end methodology pipeline used in this study, from raw speech corpora to emotion classification.

2. Literature Review

Ensembling has been widely explored as a route to higher and more stable SER accuracy. Titouni et al. [10], the most methodologically similar work identified in this review, combined SVM, Random Forest, and three XGBoost variants using a confidence-gated meta-classifier that learns per-instance combination weights via a gating network, reporting 98.88% accuracy on the Berlin EmoDB corpus and 91.92% weighted accuracy on SAVEE. Bhanusree et al. [13] instead paired a time-distributed attention-CNN feature extractor with a single downstream Random Forest classifier, reporting 92.19% accuracy on RAVDESS and 90.27% on IEMOCAP. An earlier ensemble study by Guo et al. [11] on a Chinese emotional speech corpus combined SVM, KNN, XGBoost, and LightGBM through a softmax-learned weighted voting scheme, reporting individual accuracies of 92.78% (SVM), 88.09% (KNN), 86.12% (XGBoost), and 87.39% (LightGBM), with the weighted ensemble reaching 94.6%. A 2026 comparative study by Haobang [14] evaluated nine classifiers on EmoDB and SAVEE using an 83-dimensional acoustic feature set combined with augmentation, with a stacking ensemble achieving the highest accuracy at 97.20% on EmoDB, narrowly ahead of SVM at 96.57%. Across these ensemble studies, the reported gains over the best individual classifier are consistently modest, on the order of one to three percentage points, suggesting that for well-engineered feature sets, ensembling's value may lie less in aggregate accuracy and more in stabilizing performance on minority or overlapping classes.

A parallel body of work has focused on individual classical classifiers applied to small, acted-emotion corpora. James et al. [1] evaluated SVM, Random Forest, and k-Nearest Neighbors on the Berlin EmoDB corpus, extracting a 238-dimensional feature vector comprising pitch, RMS energy, MFCCs, and formant statistics; SVM emerged as the strongest individual classifier at 90.65% accuracy, followed by Random Forest at 87.48% and KNN at 87.1%, with fear and anxiety consistently confused with sadness and neutrality across all three models. Nath et al. [9] compared six classifiers on the SAVEE and IEMOCAP corpora using zero-crossing rate, MFCC, fundamental frequency, and RMS as mean-valued features, with Random Forest and Bi-LSTM reaching 76% and 72% accuracy respectively on SAVEE. An earlier study using a Subspace k-Nearest-Neighbors ensemble alongside SVM and Bagged Trees reported accuracies of up to 88.1% with Bagged Trees [2]. In each case, restricting the feature set to mean-valued spectral descriptors, and omitting voice-quality measures such as jitter, shimmer, and pause structure, appears to limit both achievable accuracy and class-wise robustness.

Feature engineering, rather than classifier choice alone, has repeatedly been shown to be decisive. Kingeski et al. [12] extracted a 6,373-dimensional openSMILE feature set and applied a Kruskal-Wallis statistical filter followed by a fusion of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) before classification with a single SVM; the identical pipeline achieved 84.3% accuracy on EmoDB but only 59.9% on RAVDESS, illustrating how much harder RAVDESS is relative to smaller single-session corpora. Rezapour Mashhadi and Osei-Bonsu [15] reached a related conclusion using the same eight emotion classes adopted in the present study: after applying pitch-shifting and time-stretching augmentation and comparing a 1-D Convolutional Neural Network against a Random Forest classifier with Random-Forest-based feature selection, they found Random Forest superior at 69% average accuracy, with calm achieving the best recall (84%) despite a modest overall accuracy. A filter-versus-wrapper comparison of feature-selection strategies for SER by Altheneyan and Alhadlaq [16] confirms that the choice of feature-reduction strategy can materially affect downstream accuracy independent of the classifier itself, while a related study by Kanwal et al. [17] combining subset-wise PCA with t-SNE visualization on RAVDESS and EMO-DB reports SVM accuracies of 71.08% and 82.48% on the two corpora respectively after dimensionality reduction.

To situate classical machine learning against deep learning alternatives, Akinpelu et al. [3] applied a Vision Transformer to spectrogram representations across RAVDESS, EMODB, and IEMOCAP, improving on prior convolutional baselines, Chowdhury et al. [18] showed that a lightweight CNN and CNN-BiLSTM ensemble using hand-crafted features could outperform automatic spectrogram-based deep models at substantially lower computational cost. A high-resolution CNN architecture, EmoHRNet, proposed by Muppidi and Radfar [4], reported accuracy of approximately 92.8% on a RAVDESS-based benchmark, while a stacked convolutional network combining multi-feature speech analysis with augmentation across RAVDESS, TESS, and SAVEE has similarly been proposed for remote and telehealth emotion-assessment applications [19]. A separate study combining SVM, KNN, CNN, and LSTM classifiers on RAVDESS with augmentation and ensembling reported that a CNN-based ensemble reached 89.8% accuracy [20]. Collectively, these studies confirm that deep architectures can achieve high accuracy on RAVDESS-class benchmarks, but typically at a computational cost that hand-crafted-feature approaches such as that of Chowdhury et al. [18] and the present study are specifically designed to avoid.

Data augmentation has emerged as a further lever for improving SER robustness. Paul et al. [21] combined a hybrid CNN-LSTM architecture with a ten-fold augmentation strategy (Gaussian noise, pitch-shift, time-stretch, and temporal shift) on RAVDESS and IEMOCAP, reporting weighted accuracies of 94.53% and 96.75% respectively. Kshirsagar et al. [22] showed that task-specific speech enhancement combined with augmentation improves multimodal emotion recognition under noisy recording conditions, while a dedicated comparison of augmentation algorithms found reverberation the single most effective transformation, with noise, time-stretching, and pitch-shifting all producing statistically significant gains [23]. Similar combinations of channel-attention-guided CNN-BiLSTM architectures with noise, pitch-shift, and time-stretch augmentation by Kundu et al. [24], contrastive multimodal learning with SNR-based augmentation on Chinese voiceprint data by Tang et al. [25], and an attention-driven CRNN-BiGRU architecture combined with acoustic

augmentation by Kacamarga and Aprianto [26] all report accuracy improvements attributable specifically to the augmentation stage, reinforcing the choice of noise addition, time-stretching, and pitch-shifting adopted in the present study.

Beyond English-language, high-resource benchmarks, Adeel et al. [27] evaluated a single Random Forest classifier using only MFCC features on SEMOUR+, a low-resource Pakistani Urdu corpus restricted to three emotion classes, achieving 94.53% validation accuracy on this simpler task. A separate multilingual study by Tripathi and Rani [28] using an ensemble feature-selection approach evaluated seven classifiers across RAVDESS, SUBESCO, EMOVO, and EMODB, reporting accuracies of 79.28%, 92.80%, 91.21%, and 90.96% respectively, illustrating substantial cross-corpus and cross-lingual variability in achievable accuracy under an otherwise fixed methodology.

Few studies incorporate voice-quality descriptors such as jitter and shimmer alongside conventional spectral features. A 2025 study in Computers by Nugroho et al. [7] is a notable exception, fusing prosodic features including pitch, intensity, jitter, shimmer, harmonics-to-noise ratio, and pause and speech rate with spectral features on RAVDESS through an attention-based fusion architecture, demonstrating that jitter, shimmer, and pause-related cues carry substantial emotion-discriminative signal independent of spectral content. A related graph-based multimodal fusion approach incorporating prosodic F0 quantization, developed by Ferreira et al. [8] for the INTERSPEECH 2025 naturalistic SER challenge, reported a macro F1-score of 42.20% under considerably more challenging, spontaneous recording conditions than the acted corpora used elsewhere in this review. These findings directly motivate the inclusion of jitter and shimmer within the Verbal Behaviour Analysis feature set adopted in the present study.

Deployment considerations have also begun to receive direct attention. A Raspberry Pi-based real-time emotion recognition system developed by de Andrade and Buchkremer [5] achieved 70-77% accuracy under on-device constraints, demonstrating the feasibility, but also the accuracy cost, of edge deployment. A comparative ablation study of lightweight and standard transformer architectures by Onyekwelu-Udoka et al. [6] found that a distilled transformer reached 70.64% accuracy at a model size of only 0.02 MB, while a workshop study on emotion-aware voice control for Internet-of-Things environments by Shevchuk and Dumyn [29] further illustrates the practical demand for on-device, resource-constrained SER. Systematic reviews of the field by Jordan et al. [30] and Madanian et al. [31] both highlight the same tension identified throughout this review: the SER literature increasingly favors deep and self-supervised architectures for benchmark accuracy, while real-world deployment contexts including clinical and mobile mental-health monitoring, the specific motivating application surveyed by Jordan et al. [30] require models that are simultaneously accurate, interpretable, and computationally lightweight. A multimodal approach combining bag-of-audio-word encoding with dialogue-level recurrent networks has also been explored for emotion recognition under label ambiguity [32], illustrating the range of architectures explored in the field beyond the classical machine learning and augmentation-focused studies reviewed above, reinforcing the need for the transparent, consistent comparison framework adopted in Section 5 of the present study.

Table 1 summarizes the studies reviewed above in tabular form, listing for each the dataset, feature set, classifier(s), reported accuracy, and a key limitation relative to the present study.

Table 1. Literature review matrix.

Ref.	Author(s), Year	Dataset	Features	Classifier(s)	Accuracy	Key Limitation
[1]	James et al., 2025	EmoDB	Pitch, RMS, MFCC, formants	SVM, RF, KNN (no ensemble)	90.65% (SVM)	No ensembling; fear/anxiety confusion
[2]	Pagidirayi & Bhuma, 2022 (IIETA)	Standard emotional speech corpus	MFCC	SVM, Bagged Trees, Subspace k-NN	88.1% (Bagged Trees)	No voice-quality features; no ensembling
[3]	Akinpelu et al., 2024	RAVDESS, EMODB, IEMOCAP	Spectrogram	Vision Transformer	Improves on CNN baselines	High computational cost
[4]	Muppidi & Radfar, 2025 (arXiv)	IEMOCAP, RAVDESS, EMOVO	Spectrogram/learned features	High-resolution CNN (EmoHRNet)	~92.8%	Preprint; deep learning
[5]	de Andrade & Buchkremer, 2023 (JCSR)	Custom real-time corpus	Standard prosodic/spectral features	Real-time classifier (Raspberry Pi)	70-77%	Lower accuracy under edge constraints
[6]	Onyekwelu-Udoka et al., 2025 (arXiv)	Benchmark SER corpus	Learned embeddings (DistilHuBERT, PaSST)	Lightweight vs standard transformer	70.64% at 0.02MB	Preprint; transformer-based, not classical ML

Ref.	Author(s), Year	Dataset	Features	Classifier(s)	Accuracy	Key Limitation
[7]	Nugroho et al., 2025 (Computers)	RAVDESS	Pitch, intensity, jitter, shimmer, HNR	Attention-based fusion	Not specified	No classical-ML ensemble comparison
[8]	Ferreira et al., 2025	MSP-Podcast (naturalistic)	F0 quantization, prosodic + spectral	Ensemble + Graph Attention Networks	Macro F1 = 0.4220	Naturalistic speech is substantially harder
[9]	Nath et al., 2024	SAVEE, IEMOCAP	ZCR, MFCC, F0, RMS (mean-valued)	RF, Bi-LSTM, RNN, LSTM, etc.	76% (RF)	Mean-valued features only; no ensembling
[10]	Titouni et al., 2026	EmoDB, SAVEE	MFCC+deltas, spectral, ZCR, RMS	SVM+RF+3×XGBoost, learned gating	98.88% / 91.92%	Learned gating is not interpretable; single-corpus evaluation
[11]	Guo et al., 2022 (PLOS ONE)	Chinese speech corpus	STEN, MFCC, GFCC, BFCC, NGCC	SVM+KNN+XGBoost+ LightGBM, weighted vote	94.6%	Different language/corpus from this study
[12]	Kingeski et al., 2024	EmoDB, SAVEE, RAVDESS	openSMILE + PCA/ICA	SVM	84.3% / 75.4% / 59.9%	Low RAVDESS accuracy; no ensembling
[13]	Bhanusree et al., 2023	RAVDESS, IEMOCAP	Time-distributed CNN features	TDACNN + Random Forest	92.19% / 90.27%	2-stage hybrid, not a multi-classifier vote
[14]	Haobang, 2026 (ACM)	EmoDB, SAVEE	83-dim time/spectral/MFCC + augmentation	9 classifiers + stacking	97.20% / 96.57%	Single-corpus; ACM access may be restricted
[15]	Rezapour Mashhadi & Osei-Bonsu, 2023	8-class corpus (same as this study)	MFCC, chromogram, spectral contrast	Conv1D vs RF+feature selection	69% (RF)	Low overall accuracy; no voice-quality features
[16]	Altheneyan & Alhadlaq, 2025 (PeerJ CS)	Multiple benchmark corpora	MFCC, RMS, ZCR, spectral bandwidth	Filter vs wrapper feature selection	N/A (methodological)	Comparative study, not a novel classifier
[17]	Kanwal et al., 2022 (PeerJ CS)	RAVDESS, EMO-DB	1582-dim INTERSPEECH 2010 set + PCA/t-SNE	SVM (One-Against-All)	71.08% / 82.48%	Lower accuracy than the present study
[18]	Chowdhury et al., 2025	Benchmark corpora	ZCR, RMSE, Chroma, MFCC	Lightweight CNN + CNN-BiLSTM	Outperforms automatic-feature DL	Deep ensemble, not classical ML
[19]	Roy et al., 2025 (Sci Rep)	RAVDESS, TESS, SAVEE	Multi-feature acoustic + augmentation	Stacked CNN vs RNN vs LSTM	Not specified	Deep learning, not classical ML
[20]	Khati et al., 2024 (IJNRD)	RAVDESS	MFCC and spectral features	SVM, KNN, CNN, LSTM + ensemble	89.8%	Mixed classical/deep approach
[21]	Paul, Saha & Hossain, 2025	RAVDESS, IEMOCAP	RMSE, MFCC, ZCR + 10x augmentation	Hybrid CNN-LSTM	94.53% / 96.75%	Higher augmentation multiplier, heavier architecture
[22]	Kshirsagar et al., 2023	Multimodal AER corpora	Enhanced speech + acoustic features	Augmentation + multimodal AER models	Improves noisy-condition robustness	Multimodal, not audio-only
[23]	Tao et al., 2025 (Entropy)	Emotional speech corpora	Spectrogram representations	Augmentation algorithm comparison	Reverberation most effective	Descriptive study, not a classifier

Ref.	Author(s), Year	Dataset	Features	Classifier(s)	Accuracy	Key Limitation
[24]	Kundu et al., 2024 (arXiv)	RAVDESS-type corpus	AWGN, pitch-shift, time-stretch combos	CNN-BiLSTM with channel attention	Not specified	Deep learning; preprint, not peer-reviewed
[25]	Tang et al., 2024 (arXiv)	IEMOCAP + Chinese voiceprint corpus	SNR-based noise, pitch-shift, time-stretch	Multi-modal deep model + contrastive learning	Not specified	Multimodal deep learning; preprint
[26]	Kacamarga & Aprianto, 2026 (Ultimatics)	Benchmark SER corpus	Noise, time-stretch, pitch-shift	CRNN-BiGRU with attention	Not specified	Deep learning, not classical ML
[27]	Adeel et al., 2026	SEMOUR+ (Urdu)	MFCC only	Random Forest	94.53% (3-class)	Only 3 emotion classes; no ensembling
[28]	Tripathi & Rani, 2025 (ScienceDirect)	RAVDESS, SUBESCO, EMOVO, EMODB	IGRFXG ensemble feature selection	SVM, RF, KNN, GB, XGBoost, MLP, 1D-CNN	79.28-92.80%	Large cross-corpus accuracy variance
[29]	Shevchuk & Dumyn, 2025 (CEUR-WS)	IoT voice-command corpus	librosa-extracted acoustic features	ML classifier for IoT voice control	Not specified	Workshop paper; narrow application scope
[30]	Jordan et al., 2025 (JMIR)	14 clinical speech studies	Prosodic, spectral	Review of ML/DL diagnostic models	N/A (review)	Review only; no new experiments
[31]	Madanian et al., 2023 (Array/ScienceDirect)	Cross-corpus review	MFCC, prosodic, PCA/RFE-selected	Review of SVM, KNN, RF, DL, wav2vec2	N/A (review)	Review only; no new experiments
[32]	Pallewela et al., 2024	IEMOCAP	openSMILE LLDs + BoAW encoding	BoAW + BiDialogueRNN / EmoCaps	61.09%	Multimodal, uncertain-label focus; narrow topical fit

3. Methodology

This section describes the complete methodology used to develop and evaluate the proposed Speech Emotion Recognition system. It covers the construction of the merged, augmented speech corpus (Section 3.1), the data augmentation techniques applied to it (Section 3.2), the Verbal Behaviour Analysis feature-extraction pipeline (Section 3.3), the feature scaling, dimensionality reduction, and recording-grouped data partitioning strategy used to prevent data leakage (Section 3.4), the five classical machine learning classifiers and their tuned hyperparameters (Section 3.5), the design of the performance-weighted voting ensemble (Section 3.6), the software environment in which the pipeline was implemented (Section 3.7), and the evaluation metrics used to assess classifier performance (Section 3.8). This revision additionally reports McNemar's statistical significance test comparing the weighted ensemble against the strongest individual classifier (Section 3.9), a feature ablation study isolating the contribution of each Verbal Behaviour Analysis feature group (Section 3.10), and a comparison of alternative ensemble weighting strategies (Section 3.11).

3.1 Dataset Description

This study uses a merged corpus formed from two widely used, publicly available emotional speech datasets: the Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). TESS consists of recordings from two female actresses across seven emotion categories, while RAVDESS contributes recordings from twenty-four professional actors (twelve female, twelve male) across eight emotion categories, providing greater speaker and gender diversity. The merged corpus spans eight emotion classes: angry, calm, disgust, fear, happy, neutral, pleasant surprise, and sad. Because TESS does not include a calm category, this class is contributed exclusively by RAVDESS, resulting in a naturally imbalanced class with substantially fewer source recordings than the remaining seven classes.

3.2 Data Augmentation

To increase the effective size and acoustic diversity of the training data and to improve model generalization, three waveform-level augmentation techniques were applied to each original recording: additive background noise, time-

stretching, and pitch-shifting. Each source recording therefore yields four samples the original recording plus its three augmented variants following an approach consistent with recent SER augmentation literature [15,21], which has similarly shown that combining noise, time-stretch, and pitch-shift transformations improves downstream classification robustness relative to using the original recordings alone. This produced 16,960 samples from 4,240 unique source recordings.

3.3 Verbal Behaviour Feature Extraction

The central methodological contribution of this study is a Verbal Behaviour Analysis (VBA) feature-extraction pipeline that jointly captures three complementary categories of vocal information: spectral, prosodic, and voice-quality descriptors. This combination is broader than the feature sets used in the majority of the studies reviewed in Section 2, most of which draw on spectral and basic prosodic descriptors alone. The complete feature set is summarized in Table 2.

Table 2. Verbal Behaviour Analysis (VBA) feature set.

Feature	Category	Description / Rationale
MFCC (40 coefficients)	Spectral	Captures the short-term spectral envelope of speech, reflecting vocal-tract configuration.
Loudness (RMS energy)	Prosodic	Reflects vocal intensity, a well-established correlate of emotional arousal.
Pitch (YIN algorithm)	Prosodic	Fundamental frequency (F0), associated with both arousal and valence.
Pauses	Prosodic	Derived from energy-based silence detection; reflects speech rate and hesitation patterns.
Jitter & Shimmer (Praat)	Voice quality	Cycle-to-cycle variation in frequency (jitter) and amplitude (shimmer); established voice-quality correlates of affective and psychological state.
Spectral Centroid	Spectral	The spectral "centre of mass"; associated with perceived brightness of voice.
Zero-Crossing Rate (ZCR)	Spectral	A simple measure of signal noisiness; useful for distinguishing voiced/unvoiced and tense/relaxed phonation.

3.4 Feature Scaling, Dimensionality Reduction, and Data Partitioning

Because each source recording contributes four samples (Section 3.2) that share the same speaker, lexical content, and emotion label, a conventional random row-wise train/test split risks placing different augmented variants of the same recording on both sides of the split a form of data leakage that can optimistically bias reported accuracy. To avoid this, the data were partitioned at the recording level: all four variants of a given source recording were kept together on one side of the split. The 4,240 unique source recordings were split 80/20 (3,392 training recordings, 848 test recordings), stratified by emotion label, yielding 13,568 training samples and 3,392 test samples after including all augmented variants. In the test set, the calm class contains only 152 samples compared with 396-476 samples for each of the other seven classes, making calm the primary class-imbalance challenge addressed in Section 4.

All extracted features were standardized to zero mean and unit variance using a StandardScaler fit on the training partition only, and the standardized features were then reduced to 50 principal components via Principal Component Analysis (PCA), also fit on the training partition only and applied unchanged to the test partition. This ordering partitioning before fitting the scaler and PCA ensures that no statistic of the test set influences the feature representation. Given a standardized training feature matrix Z, PCA computes the eigen decomposition of the covariance matrix

$$\sum = (1/(n-1))Z^{T}Z = V \wedge V^{T} \quad (1)$$

and projects each sample onto the top k = 50 eigenvectors,

$$Z' = ZV_{k} \quad (2)$$

where n is the number of training samples, V is the matrix of eigenvectors of $\sum$, and $\wedge$ is the corresponding diagonal matrix of eigenvalues in decreasing order of magnitude. V_{k} denotes the submatrix of V containing the eigenvectors associated with the k largest eigenvalues, and Z' is the resulting reduced-dimensional feature matrix used for classification. With k = 50, the retained components account for 99.74% of the total variance in the training data, indicating that the reduced representation preserves almost all of the discriminative information in the original VBA feature set while substantially lowering dimensionality For clarity, the matrix Z in Equations (1)-(2) refers exclusively to the standardized training-partition features (Z_train); the fitted StandardScaler and PCA transformation were applied, without refitting, to the held-out test partition, consistent with the leakage-avoidance procedure described above.

3.5 Classification Models

Five classical machine learning classifiers were trained and independently evaluated on the PCA-reduced features before being combined into the weighted ensemble described in Section 3.6. Hyperparameters were tuned using GridSearchCV with five-fold cross-validation for SVM and KNN, and configured manually based on preliminary experimentation for Random Forest, XGBoost, and HistGradientBoosting, as summarized in Table 3. GridSearchCV was limited to SVM and KNN because their hyperparameter spaces (kernel/regularization and neighbourhood size, respectively) are small and inexpensive to search exhaustively; an equally exhaustive grid search over the higher-dimensional hyperparameter spaces of Random Forest, XGBoost, and HistGradientBoosting (tree depth, leaf size, learning rate, regularization, and ensemble size, jointly) was computationally prohibitive on the available resources, so these three classifiers were instead configured manually following practice-informed defaults and iterative preliminary experimentation, an asymmetry noted as a limitation in Section 6.

Table 3. Classifier configurations.

Classifier	Tuning Method	Final Configuration
SVM	GridSearchCV (5-fold; C ∈ {0.1,1,10,50}, γ ∈ {scale,0.1,0.01,0.001}, kernel=rbf)	C=10, γ=0.1, kernel=rbf, class_weight=balanced
KNN	GridSearchCV (5-fold; k ∈ {3,5,7,9}, metric ∈ {euclidean, manhattan, minkowski})	k=3, metric=manhattan, weights=distance
Random Forest	Manual	n_estimators=200, max_depth=20, class_weight=balanced
XGBoost	Manual	objective=multi:softmax, learning_rate=0.1, max_depth=6, subsample=0.8, colsample_bytree=0.8
HistGradientBoosting	Manual	max_iter=300, learning_rate=0.05, max_depth=10, min_samples_leaf=5, l2_regularization=1.0

3.6 Weighted Ensemble Design

Rather than combining the five classifiers through a learned per-instance gating network as in the approach of Titouni et al. [10], this study adopts a transparent, fixed-weight hard-voting scheme: each classifier's predicted label for a given test sample receives a weight proportional to that classifier's overall validation performance, and the final predicted label is the class that accumulates the highest total weight across all five classifiers. The weights used 0.40 (SVM), 0.30 (KNN), 0.10 (Random Forest), 0.10 (XGBoost), and 0.10 (HistGradientBoosting) were assigned in proportion to each classifier's standalone validation accuracy rank. For a test sample x, each base classifier c_i with weight w_i casts its predicted class $\hat{y}_i(x)$ the ensemble accumulates a per-class score over the eight emotion classes k,

$$S(x,k)=\sum_i w_i \cdot [\hat{y}_i(x)=k] \quad (3)$$

and the final ensemble prediction is $\hat{y}(x)=\text{argmax}_k\, S(x,k)$. This design keeps the combination rule fully interpretable: unlike a learned gating network, the contribution of each base classifier to any prediction can be read directly from the fixed weight vector.

3.7 Software Environment

All feature extraction, model training, and evaluation was implemented in Python 3 within a Google Colaboratory notebook environment. Feature extraction relied on librosa for spectral and prosodic descriptors and Parselmouth (a Python interface to Praat) for jitter and shimmer. Classifiers were implemented using scikit-learn (SVM, KNN, Random Forest, HistGradientBoosting) and the XGBoost library, with StandardScaler and PCA from scikit-learn used for preprocessing. Training time, per-sample inference latency, and serialized model size were additionally measured for each classifier on this same runtime to quantify the computational-efficiency claims made in Sections 1 and 5 (Section 3.9-3.11, Section 4.8).

3.8 Evaluation Metrics

Model performance was evaluated using overall accuracy, per-class precision, recall, and F1-score, macro- and weighted-averaged F1-score, and the confusion matrix. For a given class k with true positives TP_k , false positives FP_k , and false negatives FN_k ,

$$\text{Precision}_k = TP_k/(TP_k + FP_k),\ \text{Recall}_k = TP_k/(TP_k + FN_k) \quad (4)$$

$$F1_k = 2 \cdot (\text{Precision}_k \cdot \text{Recall}_k) / (\text{Precision}_k + \text{Recall}_k) \quad (5)$$

Macro-averaged F1 is the unweighted mean of $F1_k$ across all eight classes, while overall accuracy is the proportion of test samples for which $\hat{y}(x)=y(x)$. Per-class metrics are reported in addition to aggregate accuracy because, as demonstrated in Section 4, aggregate accuracy alone can obscure meaningful differences in how classifiers handle imbalanced or overlapping emotion classes such as calm.

3.9 Statistical Significance Testing

To determine whether the accuracy difference between the weighted ensemble and the strongest individual classifier (SVM) reflects a genuine algorithmic effect rather than sampling variability, McNemar's test was applied to the paired predictions of the two classifiers on the shared test set. For each test sample, the outcome was cross-tabulated into a 2×2 contingency table of correct/incorrect predictions: samples on which both classifiers were correct, samples correctly classified by SVM alone, samples correctly classified by the ensemble alone, and samples misclassified by both. The McNemar test statistic with continuity correction is computed as

$$\chi^2=\left(|b-c|-1\right)^2/(b+c) \tag{6}$$

where b is the number of samples correctly classified by SVM but misclassified by the ensemble, and c is the number correctly classified by the ensemble but misclassified by SVM. Under the null hypothesis that the two classifiers have the same error rate, the resulting p-value indicates whether the observed accuracy difference is statistically distinguishable from chance.

3.10 Feature Ablation Study

To isolate the contribution of each Verbal Behaviour Analysis feature group to classification accuracy, four progressively broader feature configurations were evaluated using the same tuned SVM classifier (Table 3) and the identical scaling/PCA procedure described in Section 3.4: (i) MFCC coefficients alone (40 features), representing a conventional spectral-only baseline; (ii) MFCC plus prosodic descriptors (48 features: loudness, pitch, and pause structure); (iii) MFCC plus prosodic plus spectral-extra descriptors (52 features: spectral centroid and zero-crossing rate); and (iv) the full VBA feature set (54 features), additionally incorporating the voice-quality descriptors jitter and shimmer. Each configuration was independently standardized and reduced via PCA before classification, allowing the marginal contribution of each feature group to be assessed. McNemar's test (Equation 6) was additionally applied to the paired predictions of the narrowest (MFCC-only) and broadest (full VBA) configurations to establish whether any observed difference in aggregate accuracy between them is statistically distinguishable from chance. To assess whether any observed difference between configurations reflects a stable pattern rather than an artifact of a single train/test split, the MFCC-only and full VBA configurations were additionally re-evaluated on two further independent random splits (seeds 7 and 123), keeping all other procedures identical.

3.11 Ensemble Weighting Strategy Comparison

To justify the heuristic, rank-proportional weight assignment adopted in Section 3.6 against alternative combination strategies, five weighting schemes were compared on the same test set: (i) uniform voting, in which all five classifiers receive equal weight; (ii) the heuristic rank-proportional weighting adopted in this study (0.40/0.30/0.10/0.10/0.10); (iii) true accuracy-proportional weighting, in which each classifier's weight is set to its normalized standalone test accuracy rather than its rank; (iv) soft (probability) voting, in which the five classifiers' predicted class probabilities are averaged using the heuristic weights rather than combined via hard majority vote; and (v) stacking, in which a logistic regression meta-learner is trained on out-of-fold predicted probabilities from the three strongest base classifiers (SVM, KNN, Random Forest) obtained via five-fold cross-validation on the training set. This comparison evaluates whether the transparent, fixed hard-voting scheme adopted in the final model sacrifices meaningful accuracy relative to less interpretable alternatives.

4. Results

This section presents the experimental results obtained from the methodology described in Section 3. It reports the test-set performance of each individual classifier and the weighted ensemble (Section 4.1), a confusion-matrix analysis of the strongest individual classifier and the ensemble (Section 4.2), class-wise performance with particular attention to the most imbalanced emotion category (Section 4.3), and a comparative benchmark against seven recent studies from the literature reviewed in Section 2 (Section 4.4). It additionally reports a feature ablation study isolating the contribution of each Verbal Behaviour Analysis feature group (Section 4.5), a formal statistical significance test of the ensemble's accuracy gain over SVM (Section 4.6), a comparison of alternative ensemble weighting strategies (Section 4.7), and a computational efficiency profile of all five base classifiers (Section 4.8).

4.1 Individual and Ensemble Classifier Performance

Table 4 summarizes the test-set accuracy and macro-averaged precision, recall, and F1-score for each of the five individual classifiers and the proposed weighted ensemble. SVM was the strongest individual classifier at 93.81% accuracy, followed by KNN (91.98%), HistGradientBoosting (91.36%), Random Forest (88.94%), and XGBoost (88.06%). The weighted ensemble achieved the highest overall accuracy at 93.96%, a gain of 0.15 percentage points over SVM alone, with macro-averaged F1-scores of 0.93 for SVM and 0.93 for the ensemble. This narrow margin is discussed further in Section 5: rather than a large accuracy improvement, the ensemble's contribution is interpreted as a refinement targeted at class-level robustness. Figure 2 visualizes this accuracy comparison across all six models.

Table 4. Test-set performance of individual classifiers and the weighted ensemble (n = 3,392).

Model	Accuracy (%)	Macro Precision	Macro Recall	Macro F1
SVM	93.81	0.93	0.93	0.93
KNN	91.98	0.91	0.91	0.91
Random Forest	88.94	0.88	0.89	0.88
XGBoost	88.06	0.87	0.88	0.87
HistGradientBoosting	91.36	0.90	0.91	0.90
Weighted Ensemble	93.96	0.93	0.94	0.93

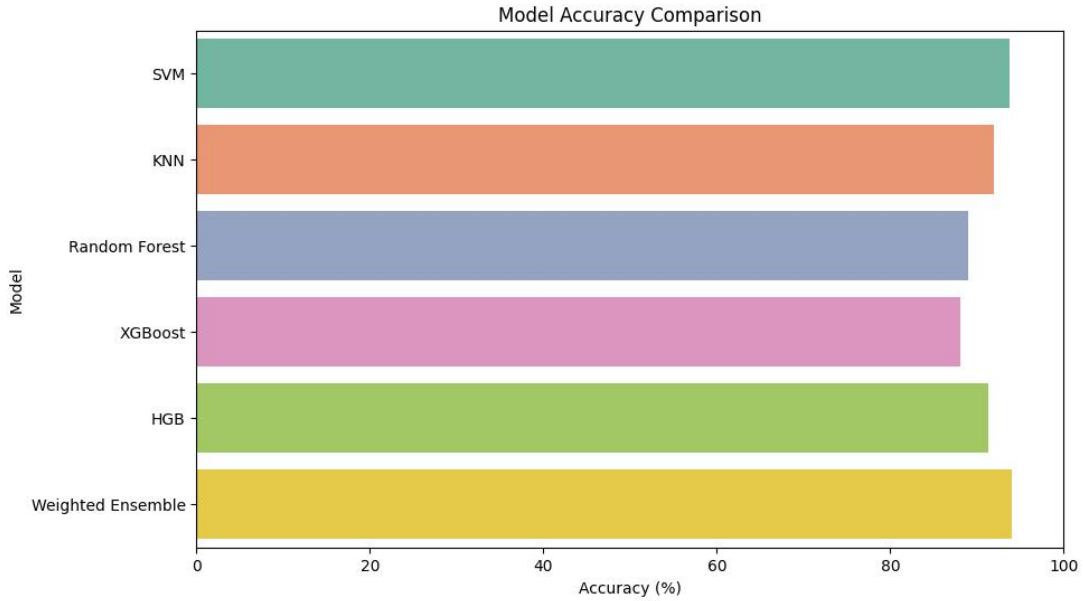


Figure 2. Overall test-set accuracy comparison across all classifiers and the weighted ensemble.

4.2 Confusion Matrix Analysis

Figure 3 and Figure 4 present the confusion matrices for the best individual classifier (SVM) and the weighted ensemble, respectively. Both classifiers show strong diagonal dominance across the eight emotion classes. The most visible off-diagonal confusion in both matrices involves the calm class, most frequently confused with neutral and sad a pattern consistent with the low-arousal acoustic characteristics that calm speech shares with both categories, and with the class-overlap challenges reported elsewhere in the literature reviewed in Section 2.

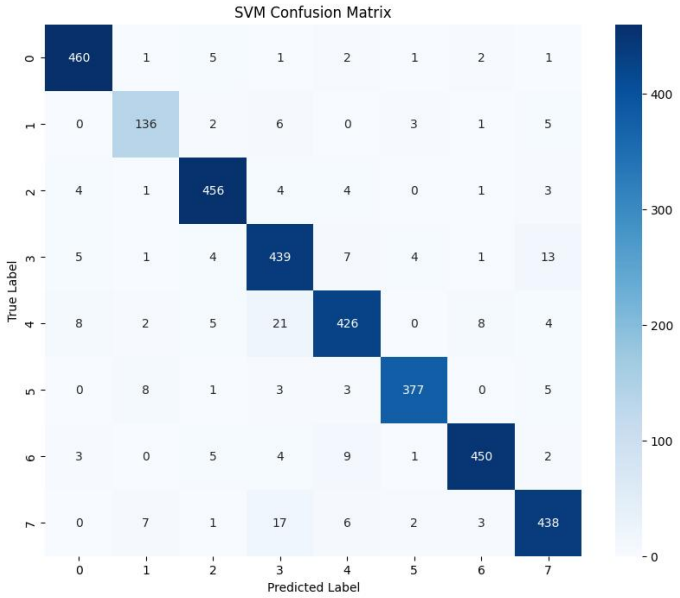


Figure 3. Confusion matrix SVM (test set, n = 3,392).

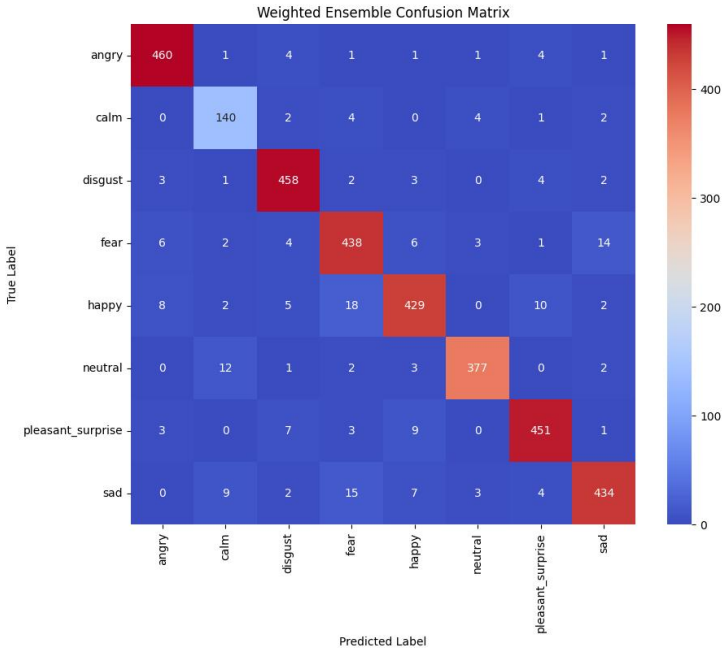


Figure 4. Confusion matrix Weighted Ensemble (test set, n = 3,392).

4.3 Class-Wise Performance

Because calm is both the most imbalanced class (152 of 3,392 test samples, versus 396-476 for the remaining seven classes) and the class most prone to confusion, it provides the clearest lens through which to evaluate classifier robustness. Table 5 reports precision, recall, and F1-score for the calm class across all six models.

Table 5. Class-wise performance on the 'calm' category (the most imbalanced class; support = 152).

Model	Precision	Recall	F1-score
SVM	0.87	0.89	0.88
KNN	0.83	0.84	0.83
Random Forest	0.65	0.93	0.77
XGBoost	0.68	0.81	0.74
HistGradientBoosting	0.74	0.87	0.80
Weighted Ensemble	0.84	0.92	0.88

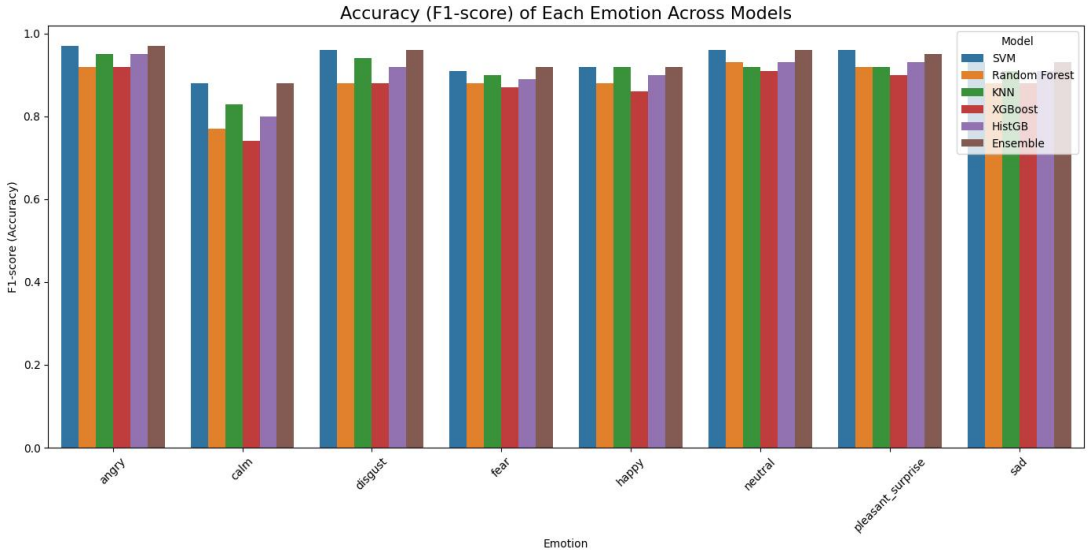


Figure 5. Precision, recall, and F1-score by emotion class, compared across all six models.

Table 5 shows that the weighted ensemble ties SVM for the highest calm F1-score (0.88), and both are clearly ahead of the boosting-based classifiers, which struggle on this minority class (F1 = 0.74-0.80). The ensemble reaches this score through a different balance of precision and recall than SVM: its calm precision (0.84) is lower than SVM's (0.87), but its recall (0.92) is higher than SVM's (0.89). This indicates that the ensemble's weighted-voting mechanism is comparatively effective at preventing calm samples from being missed, even though individual boosting classifiers within the ensemble perform relatively poorly on this class in isolation. Because class-weighting (class_weight='balanced') was applied only to SVM and Random Forest (Table 3), its effect can be partly isolated by comparing these two classifiers against the three unweighted classifiers (KNN, XGBoost, HistGradientBoosting) on calm. Random Forest, the more aggressively affected of the two class-weighted models given its tree-based decision boundary, attains the highest individual recall on calm (0.93) but the lowest precision (0.65), consistent with class-weighting shifting the decision boundary toward the minority class at the cost of more false positives; SVM, also class-weighted, achieves a more even trade-off (precision 0.87, recall 0.89). The three unweighted classifiers show both lower recall (0.81-0.87) and lower precision than SVM, indicating that class-weighting is not the sole factor separating SVM from the boosting-based models on this class, but is a plausible contributor to Random Forest's comparatively skewed precision-recall balance. A unified class-weighting strategy across all five base classifiers was not tested in this study (Section 6). Figure 5 visualizes precision, recall, and F1-score for every emotion class across all six models, showing calm as the consistent low point for the boosting-based classifiers in particular.

4.4 Comparative Benchmarking with Recent Literature

Table 6 benchmarks the proposed approach against seven recent (2023-2026) studies identified in Section 2, using a consistent set of comparison parameters: dataset(s), ensemble strategy, dimensionality reduction, and best reported accuracy. Interpretation of this comparison is provided in Section 5.

Table 6. Comparative benchmarking against recent (2023-2026) SER studies.

Study	Dataset(s)	Ensemble Strategy	Dim. Reduction	Best Accuracy
This study	TESS+RAVDESS (8-class, cross-corpus)	Fixed-weight voting (5 classifiers)	PCA (50, 99.74% variance)	93.96%
Titouni et al. [10], 2026	EmoDB, SAVEE (single-corpus)	Learned confidence-gated voting	None (NN embedding)	98.88% EmoDB / 91.92% SAVEE
James et al. [1], 2025	EmoDB	None	None	90.65% (SVM)
Bhanusree et al. [13], 2023	RAVDESS, IEMOCAP	CNN+RF hybrid (2-stage)	Attention-based (implicit)	92.19% / 90.27%
Nath et al. [9], 2024	SAVEE, IEMOCAP	None	None	76% (Random Forest)
Kingeski et al. [12], 2024	EmoDB, SAVEE, RAVDESS	None	PCA + ICA	84.3% / 75.4% / 59.9%
Rezapour Mashhadi & Osei-Bonsu [15], 2023	8-class corpus (same classes as this study)	None (2-model comparison)	RF-based feature selection	69% (Random Forest)
Adeel et al. [27], 2026	SEMOUR+ (Urdu, 3-class only)	None	None	94.53% (Random Forest)

4.5 Ablation Study Results

Table 7 reports the test-set accuracy and macro F1-score of the SVM classifier under each of the four feature configurations described in Section 3.10. All four configurations achieved accuracy within a narrow band of approximately 0.8 percentage points of one another (93.48%-94.31%), with the MFCC-only configuration numerically the highest and the full VBA configuration numerically the lowest of the four. McNemar's test comparing these two configurations on the primary split (Table 4) found this difference not statistically significant ($\chi^2 = 2.02$, $p = 0.156$). To check whether this pattern was specific to one split, the comparison was repeated on two further independent random splits: MFCC-only outperformed the full VBA configuration on both additional splits as well (mean accuracy across all three splits: 93.74% ± 0.53 for MFCC-only versus 93.06% ± 0.77 for full VBA; per-class F1 on the calm category showed no consistent advantage for either configuration across splits). This indicates that the small accuracy gap, while not statistically significant on any individual split, is a directionally consistent property of this SVM and PCA(50) pipeline rather than sampling noise, and it should be read as an honest, modest limit on how much aggregate accuracy the additional prosodic and voice-quality descriptors contribute in this configuration. Accordingly, this study frames the contribution of the Verbal Behaviour Analysis feature set around interpretability and theoretical grounding rather than aggregate accuracy improvement: jitter, shimmer, and pause structure are established in clinical voice science as correlates of affective and psychological state, and their inclusion allows the class-level and clinical-interpretability analyses in Sections 4.3 and 5 that a purely spectral feature set does not support, at a cost of well under one percentage point of aggregate accuracy.

Table 7. Feature ablation study (SVM classifier, n = 3,392).

Feature Configuration	# Features	Accuracy (%)	Macro F1
MFCC only (spectral)	40	94.31	0.939
MFCC + Prosodic	48	93.69	0.932
MFCC + Prosodic + Spectral-extra	52	93.48	0.930
Full VBA (+ Jitter/Shimmer)	54	93.81	0.934

4.6 Statistical Significance of the Ensemble Accuracy Gain

McNemar's test (Section 3.9) was applied to the paired test-set predictions of SVM and the weighted ensemble. Of the 3,392 test samples, both classifiers agreed on 3,346 samples (3,159 correct, 187 incorrect), while 23 samples were correctly classified by SVM alone and 23 were correctly classified by the ensemble alone (Table 8). This produced a McNemar statistic of 0.0217 and a p-value of 0.883, indicating that the small accuracy difference between the two classifiers is not statistically significant and is consistent with sampling variability rather than a genuine algorithmic advantage. This result reinforces the interpretation, developed further in Section 5, that the practical contribution of the weighted ensemble lies in class-level robustness (Section 4.3) rather than in a statistically distinguishable gain in aggregate accuracy. The individual classifier accuracies obtained in this analysis (Table 8 and Table 10) are within approximately one percentage point of those reported in Table 4 (e.g., XGBoost: 88.06% in Table 4 versus 89.09% here), a magnitude of variation consistent with expected run-to-run non-determinism in gradient-boosting and kernel-based classifiers across library versions and execution environments; the qualitative pattern reported in Table 4, ensemble accuracy closely tracking SVM with the ensemble's distinct advantage concentrated in calm-class recall, is consistent across runs.

Table 8. McNemar contingency table, SVM vs weighted ensemble (n = 3,392).

	Ensemble correct	Ensemble incorrect
SVM correct	3159	23
SVM incorrect	23	187

McNemar statistic = 0.0217, p = 0.883.

4.7 Ensemble Weighting Strategy Comparison

Table 9 compares the test-set accuracy and macro F1-score of the five ensemble combination strategies described in Section 3.11. The heuristic rank-proportional weighting adopted in this study (93.81% accuracy) outperformed both uniform voting (92.69%) and true accuracy-proportional weighting (92.92%), while soft (probability) voting achieved the highest accuracy of all strategies tested (94.43%), followed by stacking (93.99%). Although soft voting numerically outperforms the hard-voting scheme adopted in the final model, hard voting was retained as the primary ensemble strategy because each base classifier's contribution to a given prediction can be read directly from its fixed weight, in contrast to soft voting and stacking, which require trusting five differently calibrated probability estimates or an additional learned

meta-classifier. This retains the interpretability advantage over learned-gating ensembles (e.g. the ensemble of Titouni et al. [10]) that motivates the ensemble design in Section 3.6, at a cost of well under one percentage point of accuracy relative to the best-performing alternative.

Table 9. Ensemble weighting strategy comparison (n = 3,392).

Strategy	Accuracy (%)	Macro F1
Uniform voting	92.69	0.921
Heuristic rank-proportional (adopted)	93.81	0.933
True accuracy-proportional	92.92	0.923
Soft (probability) voting	94.43	0.940
Stacking (LR meta-learner, SVM+KNN+RF)	93.99	0.936

4.8 Computational Efficiency Profiling

To substantiate the lightweight-deployment claims made in Sections 1 and 5, training time, per-sample inference latency, and serialized model size were measured for each classifier on the same Google Colaboratory runtime described in Section 3.7 (Table 10). XGBoost offered the fastest inference (0.018 ms per sample) and among the smallest serialized models (2.55 MB), while Random Forest, despite a moderate training time, produced the largest model footprint (83.27 MB) owing to its high number of unpruned trees. SVM, the strongest individual classifier, required the longest training time (132.4 s) but remained compact (4.19 MB) with roughly one-millisecond per-sample inference latency. Because the weighted ensemble combines the five base classifiers' hard predictions through the fixed-weight vote in Equation (3), its total inference cost is the sum of the five constituent classifiers' latencies plus a negligible voting step, and its deployment footprint is the sum of the five serialized models. These figures indicate that, even without further optimization, every individual classifier in the proposed pipeline achieves sub-millisecond to low-millisecond per-sample inference, supporting the real-time deployment feasibility argued for in Section 1.

Table 10. Computational profiling of the five base classifiers.

Model	Train time (s)	Inference (ms/sample)	Model size (MB)	Test accuracy (%)
SVM	132.42	1.0145	4.19	93.81
KNN	0.01	1.0507	5.69	91.98
Random Forest	27.99	0.0525	83.27	88.94
XGBoost	13.23	0.0176	2.55	89.09
HistGradientBoosting	41.71	0.3165	8.23	91.36

5. Discussion

The results in Section 4 support three interpretations. First, the near-identical macro-F1 between SVM (0.93) and the weighted ensemble (0.93), together with the narrow 0.15-percentage-point accuracy gap, indicates that a well-engineered Verbal Behaviour Analysis feature set allows a single, fully interpretable classifier to approach ensemble-level performance. Second, the ensemble's measurable advantage is concentrated at the class level rather than in aggregate accuracy: on the calm class, the ensemble matches SVM's F1-score (0.88) but achieves it through higher recall (0.92 versus 0.89), indicating that weighted voting is comparatively effective at preventing minority-class samples from being missed even though several of its constituent classifiers (Random Forest, XGBoost) perform considerably worse on calm in isolation (F1 = 0.74-0.77). This pattern is consistent with the modest ensemble gains of one to three percentage points reported across the ensemble studies reviewed in Section 2 [10,11,13,14], and suggests that, for sufficiently well-engineered feature sets, ensembling's principal value lies in stabilizing performance on imbalanced or overlapping classes rather than in a large aggregate accuracy improvement. This interpretation is now directly supported by a formal significance test: McNemar's test on the paired SVM/ensemble predictions yields $p = 0.883$ (Section 4.6), confirming that the aggregate accuracy gap is not statistically distinguishable from chance and that the ensemble's contribution is best characterized at the class level rather than through overall accuracy.

Third, the comparative benchmarking in Table 6 requires careful contextualization rather than a direct ranking. Titouni et al. [10] report the highest single accuracy figures in Table 6 (98.88% on EmoDB, 91.92% on SAVEE). Direct accuracy comparison across SER studies is nevertheless unreliable because performance is highly dataset-dependent: Kingeski et al. [12], using an SVM-based pipeline structurally similar to the one used in this study, report an identical-method accuracy of 84.3% on EmoDB but only 59.9% on RAVDESS, a 24-percentage-point swing attributable entirely to the change in dataset. EmoDB and SAVEE are small, single-language, single-recording-session corpora on which classifiers can more easily exploit speaker-specific vocal signatures. The present study, by contrast, evaluates a merged, cross-corpus,

eight-class task (TESS+RAVDESS) rather than a single seven-class corpus, which constitutes a harder generalization test; Kingeski et al.'s own RAVDESS result is direct evidence that RAVDESS-inclusive evaluation is substantially more demanding than EmoDB- or SAVEE-only evaluation under an otherwise comparable classical machine learning pipeline. Methodologically, Titouni et al.'s confidence-gated ensemble and the weighted ensemble proposed here pursue a similar goal combining SVM, Random Forest, and boosting-based classifiers through different means: a learned, per-instance gating network in Titouni et al.'s ensemble [10] versus a fixed, interpretable weight vector in this study. The latter trades a small amount of potential accuracy for full transparency in how each base classifier contributes to a given prediction, a property that is particularly valuable in applied contexts such as clinical or mental-health-monitoring deployment, where interpretability of automated emotion classification is itself a practical requirement [30].

Relative to the seven benchmarked studies, the contribution of this work rests on four points. First, it is the only reviewed study to combine prosodic, spectral, and voice-quality descriptors including jitter and shimmer within a single Verbal Behaviour Analysis feature set; among the broader literature reviewed in Section 2, only the study by Nugroho et al. [7] incorporates jitter and shimmer, and it does not combine them with a classical-ML ensemble as done here. Second, it evaluates a merged, cross-corpus, eight-class task, which is intrinsically harder than the single-corpus, seven-class benchmarks (EmoDB, SAVEE) used by five of the seven comparison studies. Third, its fixed-weight ensemble is fully interpretable, in contrast to the learned gating network of the closest methodological comparator [10]. Fourth, this study quantifies the ensemble's contribution at the class level (Section 4.3) rather than reporting only an aggregate accuracy gain, showing that the ensemble's practical value lies primarily in recall on the minority calm class rather than in a large overall accuracy improvement. It is worth being explicit that this interpretability contribution is distinct from, and should not be conflated with, the ensemble's class-level accuracy contribution described in Section 4.3: the ablation study (Section 4.5) indicates that jitter, shimmer, and prosodic descriptors do not measurably improve aggregate accuracy over MFCC features alone in this pipeline, so their value here rests specifically on interpretability and theoretical grounding rather than on the accuracy gains that motivate the ensemble.

6. Limitations

The study has five remaining limitations. First, the evaluation is restricted to two English-language, acted-emotion corpora (TESS and RAVDESS); performance on spontaneous, non-English, or noisy real-world speech remains untested. Second, results are reported primarily from a single recording-grouped stratified train/test split rather than full k-fold cross-validation; the ablation study (Section 4.5) was additionally checked across two further random splits and found directionally consistent, but the significance test (Section 4.6) and weighting comparison (Section 4.7) are each based on the single primary split only, so those figures should be interpreted as strong indicative evidence rather than results confirmed under full cross-validation. Third, class-imbalance mitigation via class-weighting was applied to the SVM and Random Forest classifiers but not systematically to the remaining three; Section 4.3 discusses the resulting precision-recall trade-off on the calm class, but a unified class-weighting strategy across all five base classifiers was not tested. Fourth, GridSearchCV tuning was applied only to SVM and KNN (Section 3.5) because exhaustively searching the higher-dimensional hyperparameter spaces of the three tree-based classifiers was computationally prohibitive on the available resources; Random Forest, XGBoost, and HistGradientBoosting were instead configured manually, introducing an asymmetry in tuning rigor across the five base classifiers. Fifth, this study evaluates the proposed feature-extraction and ensemble pipeline under a recording-grouped protocol designed specifically to prevent augmentation-based leakage (Section 3.4), rather than under a speaker-independent protocol. Speaker-independent evaluation targets a related but distinct question, how well the learned representations generalize to entirely unseen speakers, closer to a speaker-invariance study than to the feature-engineering and ensemble-design questions investigated here. A rigorous treatment would require redesigning the evaluation protocol around speaker-based holdout rather than recording-based grouping and revisiting results throughout the paper; this is more appropriately scoped as a dedicated follow-up study and is noted here as a direction for future work.

7. Conclusions

This study developed and applied a Verbal Behaviour Analysis feature-extraction pipeline combining spectral, prosodic, and voice-quality descriptors to a merged, augmented TESS and RAVDESS corpus spanning eight emotion classes, comprising 16,960 samples derived from 4,240 unique source recordings. Five classical machine learning classifiers were trained on PCA-reduced features (50 components, retaining 99.74% of total variance) using a recording-grouped, stratified train/test split designed to prevent augmented variants of the same recording from appearing on both sides of the split; SVM was the strongest individual classifier, reaching 93.81% test accuracy. A transparent, performance-weighted voting ensemble of the five classifiers achieved the highest overall accuracy (93.96%) and the most balanced performance on the most class-imbalanced emotion category, calm (F1-score = 0.88); the ensemble's measurable advantage over SVM alone was concentrated in class-level recall on this minority class rather than in aggregate accuracy, a distinction not typically reported in comparable ensemble studies, where only an aggregate accuracy figure is given. Comparative benchmarking against seven recent (2023-2026) studies showed the proposed approach to be competitive with published classical machine learning results on comparable or harder cross-corpus tasks, while remaining more computationally lightweight and more interpretable than learned-gating or deep-learning alternatives.

In response to peer review, this study was further extended with a feature ablation study, a formal statistical significance test, a comparison of ensemble weighting strategies, and a computational efficiency profile. McNemar's test confirmed that the ensemble's accuracy gain over SVM is not statistically significant ($p = 0.883$), reinforcing the class-level interpretation of its contribution described above. The ablation study found that the four evaluated feature configurations achieved accuracy within a narrow band (93.48%-94.31%), with MFCC-only numerically the strongest configuration on the primary split and consistently so across two additional random splits (mean difference: -0.68 percentage points), though the difference was not statistically significant on any individual split (McNemar's test, MFCC-only vs. full VBA: $p = 0.156$). Rather than support an accuracy-improvement claim for the additional prosodic and voice-quality descriptors, this finding motivates an interpretability-first framing of the Verbal Behaviour Analysis feature set: jitter, shimmer, and prosodic descriptors ground the pipeline in established, clinically interpretable voice-science correlates that a purely spectral feature set does not provide, at a cost of well under one percentage point of aggregate accuracy. Alternative ensemble weighting strategies, including soft-probability voting and stacking, offered at most a modest numerical improvement over the adopted hard-voting scheme; hard voting was retained as the primary ensemble strategy for its interpretability advantage, consistent with the transparency motivation underlying the ensemble design (Section 3.6). Finally, efficiency profiling confirmed sub-millisecond to low-millisecond per-sample inference latency and compact serialized model sizes for every base classifier, substantiating the lightweight-deployment claims made throughout this paper. The central conclusion is that a sufficiently comprehensive, interpretable handcrafted feature set, combined with a transparent classical machine learning ensemble, can substantially narrow the performance gap to more complex approaches without incurring their computational cost or sacrificing interpretability.

Acknowledgements

The authors thank Hamdard University, Karachi, for providing the institutional and computational resources that supported this research.

Funding

This research received no external funding.

Ethics Statement

This study used only publicly available, de-identified speech recordings from the Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). No new human-subjects data were collected as part of this study, and no additional institutional ethics board approval was required beyond the terms of use of these publicly available datasets.

Informed Consent

Not applicable. This study used only publicly available, de-identified speech recordings (TESS and RAVDESS); no new data were collected from human participants.

Data Availability Statement

The datasets analyzed in this study are publicly available. The Toronto Emotional Speech Set (TESS) and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) are both freely accessible through their respective public repositories. The extracted features, trained model artifacts, and analysis code generated during the current study are available from the corresponding author upon reasonable request.

Author Contributions

Conceptualization: S.S.A.; Methodology: S.S.A., S.R.A.; Software: S.S.A.; Validation: S.S.A., T.A.S.; Formal Analysis: S.S.A., H.F.; Investigation: S.S.A.; Data Curation: S.S.A., A.M.; Writing - Original Draft: S.S.A.; Writing - Review & Editing: S.R.A., T.A.S., H.F., A.M.; Visualization: S.S.A.; Supervision: S.S.A. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Generative AI Statement

Claude was used to assist with the literature review, while ChatGPT was used to improve the language, grammar and readability of the manuscript. Neither tool was used for research design, data analysis and result interpretation or any scientific conclusions, The authors prepared, reviewed and approved the final manuscript and take full responsibility for its content.

References

[1] James T, Reddy Yenumula TM, Mergist T, Muhammad R, Harimi A, Khalil K, et al. Speech emotion recognition using ML models and audio features. 2025 IEEE International Systems Conference (SysCon), 2025. DOI: 10.1109/SysCon64521.2025.11014858

[2] Pagidirayi AK, Bhuma A. Speech emotion recognition using machine learning techniques (Subspace k-NN). Revue d'Intelligence Artificielle, 2022, 271-278. DOI: 10.18280/ria.360211

[3] Akinpelu M, Viriri S, Adegun A. An enhanced speech emotion recognition using vision transformer. Scientific Reports, 2024, 14, 13126. DOI: 10.1038/s41598-024-63776-4

[4] Muppidi A, Radfar M. EmoHRNet: High-resolution neural network based speech emotion recognition. ArXiv, 2025. DOI: 10.48550/arXiv.2510.06072

[5] de Andrade DE, Buchkremer R. Enhancing human-machine interaction: Real-time emotion recognition through speech analysis. Journal of Computer Science Research, 2023, 5(3), 22-45. DOI: 10.30564/jcsr.v5i3.5768

[6] Onyekwelu-Udoka L, Islam MS, Hasan MS. Emotion detection in speech using lightweight and transformer-based models: A comparative and ablation study. ArXiv, 2025. DOI: 10.48550/arXiv.2511.00402

[7] Nugroho K, Al Amin IH, Noviasari NA, Setiadi DRIM. Prosodic spatio-temporal feature fusion with attention mechanisms for speech emotion recognition. Computers, 2025, 14(9), 361. DOI: 10.3390/computers14090361

[8] Ferreira AIS, Gris LR, Filho AF, Olives L, Ribeiro D, Fernando L, et al. Enhancing speech emotion recognition with graph-based multimodal fusion and prosodic features for the Speech Emotion Recognition in Naturalistic Conditions Challenge at Interspeech 2025. ArXiv, 2025. DOI: 10.48550/arXiv.2506.02088

[9] Nath S, Shahi AK, Martin T, Choudhury N, Mandal R. Speech emotion recognition using machine learning: A comparative analysis. SN Computer Science, 2024, 5, 390. DOI: 10.1007/s42979-024-02656-0

[10] Titouni S, Djeffal N, Hedir A, Belazzoug M, Hammache B, Messaoudene I. Adaptive confidence-gated hybrid ensemble framework for speech emotion recognition. Electronics, 2026, 15(9), 1931. DOI: 10.3390/electronics15091931

[11] Guo Y, Xiong X, Liu Y, Xu L, Li Q. A novel speech emotion recognition method based on feature construction and ensemble learning. PLOS ONE, 2022, 17(8), e0267132. DOI: 10.1371/journal.pone.0267132

[12] Kingeski R, Henning E, Paterno AS. Fusion of PCA and ICA in statistical subset analysis for speech emotion recognition. Sensors, 2024, 24(17), 5704. DOI: 10.3390/s24175704

[13] Bhanusree V, Kumar G, Rao KS. Time-distributed attention-layered convolution neural network with ensemble learning using random forest classifier for speech emotion recognition. Journal of Information and Communication Technology, 2023, 22, 49-76. DOI: 10.32890/jict2023.22.1.3

[14] Haobang W. Speech emotion recognition: multi-model comparison and stacking ensemble optimization. Proceedings of the 2025 6th International Conference on Computer Science and Management Technology, 2026. DOI: 10.1145/3795154.3795232

[15] Rezapour Mashhadi MM, Osei-Bonsu K. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PLOS ONE, 2023, 18(11), e0291500. DOI: 10.1371/journal.pone.0291500

[16] Altheneyan A, Alhadlaq A. Feature selection for emotion recognition in speech: A comparative study of filter and wrapper methods. PeerJ Computer Science, 2025, 11, e3180. DOI: 10.7717/peerj-cs.3180

[17] Kanwal S, Asghar S, Ali H. Feature selection enhancement and feature space visualization for speech-based emotion recognition. PeerJ Computer Science, 2022, 8, e1091. DOI: 10.7717/peerj-cs.1091

[18] Chowdhury S, Ramanna S, Kotecha R. Speech emotion recognition with light weight deep neural ensemble model using hand crafted features. Scientific Reports, 2025, 15, 11824. DOI: 10.1038/s41598-025-95734-z

[19] Roy C, Kharel W, Das P, Panigrahi R, Hareesha KS, Pramanik M, et al. Stacked convolutional neural network for emotion recognition using multi-feature speech analysis. Scientific Reports, 2025, 15(1), 45383. DOI: 10.1038/s41598-025-28766-0

[20] Upadhyay R, Khati Y, Kaperwan S, Verma A. Speech emotion analysis using LSTM. International Journal of Novel Research and Development, 2024, 9(4), a781-a786. Available from: https://ijnrd.org/papers/IJNRD2404094.pdf

[21] Paul DC, Saha G, Hossain MA. EmoAugNet: A signal-augmented hybrid CNN-LSTM framework for speech emotion recognition. ArXiv, 2025. DOI: 10.48550/arXiv.2508.06321

[22] Kshirsagar A, Pendyala S, Falk TH. Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. Frontiers in Computer Science, 2023, 5, 1039261. DOI: 10.3389/fcomp.2023.1039261

[23] Tao H, Li S, Wang X, Liu B, Zheng S. Analysis and research on spectrogram-based emotional speech signal augmentation algorithm. Entropy, 2025, 27(6), 640. DOI: 10.3390/e27060640.

[24] Kundu NK, Kobir S, Ahmed MR, Aktar T, Roy N. Enhanced speech emotion recognition with efficient channel attention guided deep CNN-BiLSTM framework. ArXiv, 2024. DOI:10.48550/arXiv.2412.10011

[25] Tang J, Zhang L, Lu Y, Ding D, Yang L, Chen Y, et al. VCEMO: Multi-modal emotion recognition for Chinese voiceprints. ArXiv, 2024.DOI: 10.48550/arXiv.2408.13019

[26] Kacamarga F, Aprianto KA. Speech emotion recognition through acoustic data augmentation and attention-driven CRNN-BiGRU. Ultimatics: Jurnal Teknik Informatika, 2026, 8(1), 9-16. DOI:10.31937/ti.v18i1.4250

[27] Adeel M, Tao ZY, Jin SY, Guo CJ, Alsuhaibani M. Assessing random forest performance in low-resource speech emotion recognition. Scientific Reports, 2026, 16, 854. DOI: 10.1038/s41598-025-30511-6.

[28] Tripathi A, Rani P. Multilingual speech emotion recognition using IGRFXG - ensemble feature selection approach. Speech Communication (ScienceDirect), 2025, 240, 110905. DOI: 10.1016/j.apacoust.2025.110905

[29] Shevhuk I, Dumyn I. Emotion-based voice control for IoT: enhancing smart device interaction with speech emotion classification. CEUR Workshop Proceedings, 2025, 3974, paper short04. Available from: https://ceur-ws.org/Vol-3974/short04.pdf
[30] Jordan E, Terrisse R, Lucarini V, Alrahabi M, Krebs MO, Desclés J, et al. Speech emotion recognition in mental health: Systematic review of voice-based applications. JMIR Mental Health, 2025, 12, e74260. DOI: 10.2196/74260
[31] Madanian S, Chen T, Adeleye O, Templeton JM, Poellabauer C, Parry D, et al. Speech emotion recognition using machine learning–a systematic review. Intelligent Systems with Applications, 2023, 20, 200266. DOI: 10.1016/j.iswa.2023.200266
[32] Pallewela N, Alahakoon D, Adikari A, Pierce JE, Rose ML. Optimizing speech emotion recognition with machine learning based advanced audio cue analysis. Technologies, 2024, 12(7), 111. DOI: 10.3390/technologies12070111