AD EDUXIAN JOURNAL
A Quarterly Multidisciplinary Blind Peer Reviewed & Refereed Online International Journal
Google Scholar Indexing PIF:1.048(I2OR) & 3.125 (IIFS)
Vol (2), Issue (3), Oct-Dec 2025 *ISSN: 3048-7951*

Ensemble-Based Emotional Speaker Identification

Debasis Mohanta[1] and Jainath Yadav[2]

https://doi.org/10.5281/zenodo.17993927

Review: 09 /11/2025 **Acceptance: 12/11/2025** **Publication: 20/12/20251**

Abstract: This study presents an ensemble-based framework for speaker identification us- ing MFCC features extracted from an emotional speech corpus. Speaker identification is performed separately for each emotion as well as on the combined dataset to examine how emotional variability influences speaker discriminative acoustic cues. Three classical classifiers, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forests (RF), are integrated through a meta-classifier-based decision fusion strategy, where an SVM meta-learner combines the complementary decision boundaries learned by the base models. By aggregating classifier decisions rather than raw feature representations, the proposed fusion mechanism enhances robustness against emotional variations and strengthens class separation in the speaker space. The system is evaluated using accuracy, weighted and macro- averaged precision, recall, F1-scores, and confusion matrices, providing a comprehensive assessment of model behavior under different emotional conditions. The fusion framework demonstrates strong performance, achieving an accuracy of 96.17%, highlighting its effectiveness in capturing reliable speaker-discriminative patterns across emotional contexts. Further analysis using the Friedman and Nemenyi post- hoc tests statistically validates the significance of performance differences among the individual classifiers and the fused ensemble, confirming the superiority of the proposed decision-fusion approach for emotion-resilient speaker identification.

Keywords: Emotional; Speaker; Environment; Ensemble; Fusion;

Introduction: Speaker identification refers to recognizing an individual solely from the unique acoustic patterns present in their speech signal [1]. Foundational work in the field demonstrated that time–frequency–energy pattern matching and spectral cross-correlation could effectively discriminate between speakers, forming the basis for contemporary speaker- recognition technologies. Over the years, research has advanced significantly; however, identifying speakers in emotionally rich environments continues to pose substantial challenges. Emotional variations alter vocal tract behavior, prosodic characteristics, and spectral distributions, thereby disturbing speaker-specific signatures and reducing sys- tem reliability. Consequently, designing emotion-resilient speaker identification models remains an important and active area of investigation.

To address these challenges, this study presents an ensemble-based speaker identification framework leveraging MFCC [2–4] features extracted from an emotional speech cor- pus. Speaker identity is analyzed both within individual emotions and across a combined emotional dataset to understand how emotional variability influences discriminative cues. Three classical classifiers, Support Vector Machines (SVM) [2, 5–8], K-Nearest Neigh- bors (KNN), and Random Forests (RF), are integrated through a met7a-classifier-driven decision-level fusion strategy, where an SVM meta-learner aggregates complementary decision boundaries learned by the base models. This

[1]Department of Computer Science, Central University of South Bihar, Gaya, Bihar, India.
[2]Department of Computer Science, Central University of South Bihar, Gaya, Bihar, India

fusion mechanism enhances robust- ness against emotional shifts, improves class separability, and mitigates inconsistencies caused by affective fluctuations. Comprehensive evaluation using accuracy, weighted and macro precision, recall, F1-scores, and confusion matrices demonstrates that the proposed system performs consistently across emotional conditions, achieving a peak accuracy of 96.17%. Furthermore, statistical analyses using the Friedman and Nemenyi post-hoc tests validate the significance of performance differences among the classifiers, confirming the superiority of the decision-fusion approach for robust and emotion-resilient speaker identification. The major contributions of this work are summarized as follows:

- We propose an ensemble-based speaker identification framework that integrates SVM, KNN, and RF classifiers through an SVM meta-classifier for robust decision- level fusion.
- We conduct both emotion-wise and combined evaluations to investigate the impact of emotional variability on speaker-discriminative acoustic cues.
- We enhance robustness against emotional variations by fusing complementary deci- sion boundaries learned by the base classifiers, resulting in stronger class separation.
- We provide a comprehensive performance assessment using precision, recall, F1- scores, accuracy, and confusion matrices to thoroughly analyze model behavior under different emotional conditions.
- We perform statistical significance analysis using the Friedman and Nemenyi post- hoc tests, validating the superiority of the proposed decision-fusion approach.

The remaining sections are organized as follows: Section 2 presents the related surveys covering prior research on speaker identification and fusion strategies. Section 3 describes the proposed methodology, including feature extraction, classifier design, and the ensemble fusion framework. Section 4 reports the experimental results along with detailed discussion and analysis under various emotional conditions. Finally, Section 5 concludes the work and highlights key findings and potential future directions.

2. Related Survey: The field of automatic speaker recognition traces its origins back to Pruzansky's 1963 study, which first demonstrated the sufficiency of spectral information alone for high recognition performance using a simple pattern-matching procedure on time-frequency- energy patterns [1]. Significant advancement was marked by 1994, when the introduction of Gaussian Mixture Model (GMM) classifiers and cepstral mean removal were established as crucial techniques providing major gains for robust speaker identification [9]. Modern Automatic Speaker Recognition (ASR) systems typically employ Linear Prediction for ini- tial feature extraction, constructing speaker models through both unsupervised classifiers (like GMMs) and supervised methods, with contemporary efforts increasingly focusing on data fusion techniques to enhance robustness [5]. The power of hybrid approaches was quickly recognized, notably in 2001 with the hybrid GMM/Support Vector Machine (SVM) approach, which achieved an up to 25% relative reduction in identification error rate by integrating the robustness of generative models with the discriminative power of SVMs [6]. Further leveraging multi-model strength, Data fusion combining the Hidden Markov Model (HMM), Nonlinear Trajectory Normalization (NTN), and Dynamic Time Warping (DTW) models achieved an exceptional robustness, yielding an Equal Error Rate (EER) as low as 0.03% on the Multimedia database [10]. Innovations in kernel design, such

as the SVM approach utilizing Probabilistic Distance Kernels (SVM AHS) derived from GMMs, achieved a high speaker identification accuracy of 79.7% on the smaller KING corpus in 2003 [7]. By 2009, the practical hybrid GMM-SVM system em- ployed a two-stage testing process, leading to an accuracy increase from a 70.1% GMM baseline to 72.4% on the NTIMIT corpus [8]. A comprehensive literature review covering 2011 to 2016 solidified the status of Mel-Frequency Cepstral Coefficients (MFCCs) as the dominant and most successful feature extraction method, being used in 97% of reported publications [3]. More recently, the focus has shifted toward complex neural and ensemble architectures: the Multimodal Neural Network (MNN) system utilizing Wavelet Packet Transform (WPT) features and parallel neural networks achieved a high identification rate of 97.5% on the GRID database in 2015 [11]; simultaneously, a hybrid SVM/HMM system for the Oriya language achieved an enhanced accuracy of 75% by integrating a speech recognition task [2]. The pursuit of superior generalization led to the strong hybrid Random Forest (RF)-AdaBoost classification algorithm in 2021, which achieved a peak accuracy of 98.53% by addressing multi-class imbalanced speaker data [12]. Cutting-edge techniques include the Relation-based Attentive Correction Prototype Network (RACP) for few-shot recognition, which achieved a leading average accuracy of 98.11% in a 5- way 5-shot scenario [13], and a system for simultaneous identification and localization using feature fusion and a Restricted Boltzmann Machine (RBM), yielding a top accu- racy of 99.84% [14]. Finally, an analysis of classification efficiency in 2024 highlighted the k-Nearest Neighbour (k-NN) classifier's superior identification rate of 94.45% over the SVM classifier's 92.90% when using Principal Component Analysis (PCA) for feature extraction [15].

The contemporary landscape of Automatic Speaker Identification (ASR) is dominated by Deep Learning (DL) architectures, succeeding classical methods and achieving near- perfect accuracy. A notable system from 2018 achieved 97.91% accuracy by combining MFCC and UMRT features with an Artificial Neural Network (ANN), simultaneously reducing system complexity [16]. The shift toward fixed, robust speaker representations was underscored in 2019 by a novel text-independent system that used structured self- attention with deep Convolutional Neural Networks (CNNs) like VGG/ResNets to extract fixed speaker embeddings, significantly outperforming traditional i-vector methods on the VoxCeleb database [17]. Reflecting this success, a comprehensive review in 2021 surveyed the field, detailing that many successful implementations frequently achieve recognition accuracies in the 98% to 100% range using both handcrafted and deep learning features with various Machine Learning (ML) and DL classifiers [18]. Further system sophistica- tion was shown in 2022 by a novel system that utilized a pre-trained Deep Neural Network (DNN) mask for learned voice segregation (incorporating a WPT filter-bank) before clas- sification with Speech VGG, achieving superior performance with average identification rates up to 87.0% on the SUSAS dataset [19]. Addressing affective aspects of speech, an emotional speaker identification system using a modified CapsNet-M architecture with MFCCs achieved up to 89.85% average accuracy for short utterances, training faster than baseline CNNs [20]. Robustness against environmental factors remains a key focus: a 2023 study addressed noise by training state-of-the-art networks (CNN and SincNet) on novel non-speaker embeddings (silence and noise) alongside speaker classes, leading to a significantly reduced Classification Error Rate (CER), with SincNet achieving the best performance at 0.8% [21]. Another strategy for enhanced performance, demonstrated in 2024, leveraged Multiple Active Voice Detection (AVD) techniques—specifically combin- ing Zero Crossing Rate (ZCR) and Short-Term Energy (STE)—before

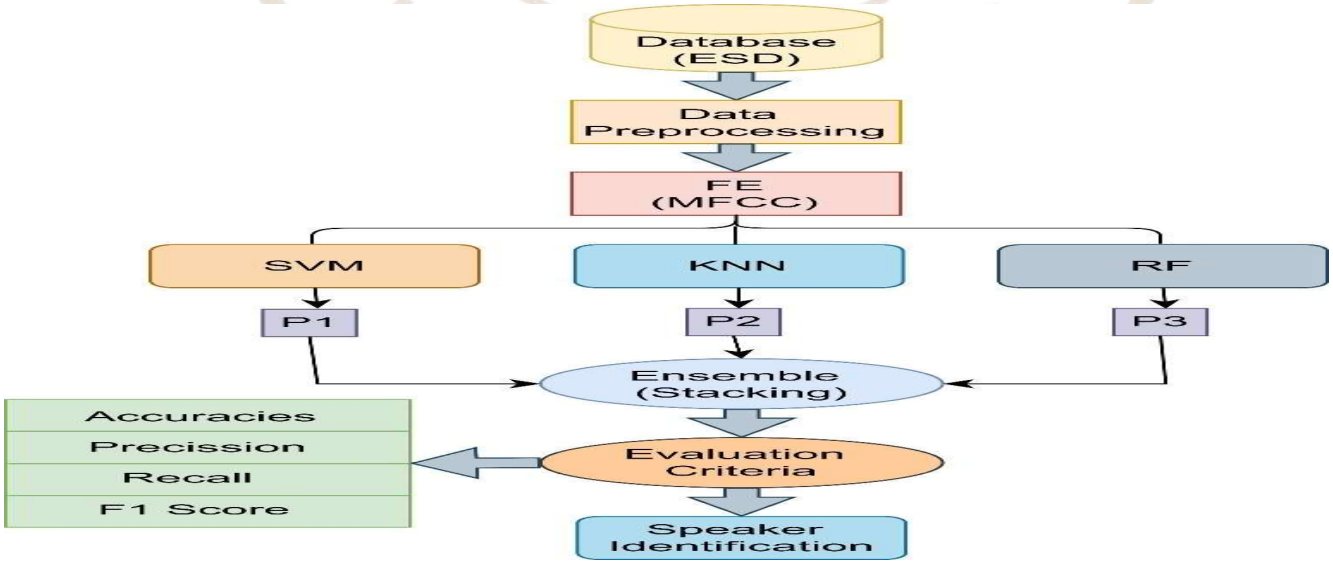
extracting MFCC features and classifying them with a Feedforward Neural Network (FFNN), which resulted in an improved accuracy of 89.25% (for 10 speakers) and performed almost 5% higher than using original signal features alone [4]. Finally, a robust emotion embedding frame- work was proposed in 2024 to address emotional inefficiency by using pre-trained DNNs to extract and extend emotional embeddings and an emotional self-attention mechanism to weight them, achieving state-of-the-art results including an Identification Rate (IR) of 59.14% on the MASC corpus and 75.98% IR on the CREMA-D corpus [22].

Methodology: This section outlines the complete pipeline of the proposed speaker identification frame- work along with dataset details. MFCC features are extracted from emotional speech data, and both emotion-specific and combined datasets are generated. The three classi- fiers, SVM, KNN, and RF, are individually trained to capture speaker distinctive char- acteristics. Their predictions are subsequently integrated through an SVM-based meta- classifier, enhancing overall robustness by leveraging complementary decision patterns.

Dataset Details: The ESD (Emotional Speech Database) is a large, high-quality emotional speech corpus designed specifically for emotional voice conversion and speech synthesis research. It con- tains recordings from 20 speakers—10 native English and 10 native Chinese—balanced by gender and aged between 25–35, all speaking in controlled studio environments with an SNR above 20 dB and a 16 kHz sampling rate. Each speaker contributes 350 par- allel utterances for five emotions (Neutral, Happy, Angry, Sad, Surprise), totaling 1750 utterances per speaker and 29 hours of speech overall [23].

Figure 1 illustrates the complete workflow of the proposed ensemble-based speaker identification system. The process begins with the ESD emotional speech database, fol- lowed by data preprocessing to prepare the audio signals for analysis. MFCC features are then extracted and fed into three individual classifiers, SVM, KNN, and RF, which independently generate prediction outputs (P1, P2, and P3). These outputs are com- bined through a stacking-based ensemble model that learns a more robust final decision by leveraging the strengths of all three classifiers. The ensemble's performance is evalu- ated using standard metrics such as accuracy, precision, recall, and F1-score, ultimately leading to the final speaker identification outcome. A detailed paradigm of the proposed framework is presented in Algorithm 1.

Figure 1: Proposed Speaker Identification Framework



Algorithm 1 Speaker Identification Framework with Decision-level Stacking

Require: Emotional speech dataset $D = \{(x_,, y_,)\}$, pre-emphasis factor α, silence threshold T, sampling rate f_s

Ensure: *Trained stacking model and evaluation metrics*

Step 1: Preprocessing

Load audio $x(t)$ at f_s and **normalize:**

$$x_{norm}(t) = \frac{x(t)}{\max\ [x(t)]}$$

Remove silence using dB threshold:

$$(x_{norm}(t),\ 10\ \log_{10}((x_{nom}(t))^2/_e > T$$

Apply pre-emphasis:

$$x_{pre}(t) = x_{active}(t) - \alpha x_{active}(t-1)$$

Step 2: Feature Extraction

Extract MFCC feature vector F_{MFCC} for each processed signal.

Step 3: Train/Test Split

Divide data into D_{train} and D_{test}.

Step 4: Base Classifier Training

Train SVM, KNN, and RF classifiers:

$$h_1 = \text{SVM}, \quad h_2 = \text{KNN}, \quad h_3 = \text{RF}$$

Step 5: Meta-Feature Generation

for each sample x do

$$p_1(x) = h_1(x), \quad p_2(x) = h_2(x), \quad p_3(x) = h_3(x)$$
$$m(x) = [p_1(x), p_2(x), p_3(x)]$$

end for

Step 6: Meta-Classifier Training

$g \leftarrow \text{SVM}(M_{train}, y_{train})$

Step 7: Prediction and Evaluation

$\hat{y} = g(m(x))$

Compute Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

Dataset Details: This section presents the experimental results obtained across all emotional conditions, comparing the performances of SVM, KNN, RF, and the proposed stacking ensemble. Metrics such as confusion matrices, precision, recall, and F1-scores are analyzed to un- derstand classifier behavior and error patterns. The Friedman test, followed by the Ne- menyi post-hoc analysis, is used to statistically validate performance differences among the models. Overall, the stacking framework consistently outperforms individual classi- fiers, demonstrating improved robustness and generalization under emotional variability.

Tables 1, 2, 3, and 4 summarizing the Classification Performance by Emotion for four distinct models—Stacking Ensemble, Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Random Forest (RF)—clearly establish the Stacking Ensemble as the best performer. The Stacking model achieved the highest overall average accuracy of 0.9617 across all emotions, just surpassing the SVM, which was the next best classifier with an average accuracy of 0.9613. The remaining models trailed significantly, with k-NN recording 0.9477 and Random Forest, the weakest performer, achieving 0.9289 average accuracy. This superiority of the Stacking Ensemble is confirmed by its highest individual emotion accuracy, classifying the Sad emotion with 0.9714 accuracy. Consistent across all classification algorithms, the Sad emotion proved the most distinct and identifiable, while the Angry emotion generally yielded the lowest performance metrics for the non-ensemble models.

Table 5 compares the accuracy of different classifiers across five emotional categories Angry, Happy, Neutral, Sad, and Surprise along with their overall averages. Among the baseline models, SVM performs the best, followed by KNN, while RF shows the low- est accuracy across emotions. The proposed stacking model achieves the highest overall accuracy (0.9617) and maintains consistently strong performance in every emotional con- dition, slightly surpassing SVM. Overall, the results highlight that the stacking approach effectively combines the strengths of individual classifiers to deliver more reliable speaker identification in emotionally varied speech.

Table 1. SVM Classification Performance by Emotion (MFCC)

W = Weighted, M = Macro, Prec = Precision, Rec = Recall, F1 = F1-score

Emotion	Acc	W-Prec	W-Rec	W-F1	M-Prec	M-Rec	M-F1
Angry	0.9486	0.9486	0.9492	0.9486	0.9486	0.9492	0.9486
Happy	0.9636	0.9635	0.9642	0.9636	0.9635	0.9642	0.9636
Neutral	0.9657	0.9657	0.9662	0.9657	0.9657	0.9662	0.9657
Sad	0.9707	0.9706	0.971	0.9707	0.9706	0.971	0.9707
Surprise	0.9579	0.9578	0.958	0.9579	0.9578	0.958	0.9579
Average	0.9613	0.9612	0.9617	0.9613	0.9612	0.9617	0.9613

Table 2. k-NN Classification Performance by Emotion (MFCC)

KNN = k-Nearest Neighbors; other abbreviations same as Table 1

Emotion	Acc	W-Prec	W-Rec	W-F1	M-Prec	M-Rec	M-F1
Angry	0.9343	0.9341	0.9349	0.9343	0.9341	0.9349	0.9343
Happy	0.9443	0.944	0.945	0.9443	0.944	0.945	0.9443
Neutral	0.9486	0.9482	0.9495	0.9486	0.9482	0.9495	0.9486
Sad	0.9679	0.9677	0.9682	0.9679	0.9677	0.9682	0.9679
Surprise	0.9436	0.9434	0.9445	0.9436	0.9434	0.9445	0.9436
Average	0.9477	0.9475	0.9484	0.9477	0.9475	0.9484	0.9477

Table 3. Random Forest Classification Performance by Emotion (MFCC)

RF = Random Forest; other abbreviations same as Table 1

Emotion	Acc	W-Prec	W-Rec	W-F1	M-Prec	M-Rec	M-F1
Angry	0.9193	0.9193	0.92	0.9193	0.9193	0.92	0.9193
Happy	0.925	0.9251	0.926	0.925	0.9251	0.926	0.925
Neutral	0.9336	0.9333	0.9337	0.9336	0.9333	0.9337	0.9336
Sad	0.9421	0.9419	0.9424	0.9421	0.9419	0.9424	0.9421
Surprise	0.9243	0.9242	0.9248	0.9243	0.9242	0.9248	0.9243
Average	0.9289	0.9288	0.9294	0.9289	0.9288	0.9294	0.9289

Table 4. Stacking Classification Performance by Emotion (MFCC)

Proposed stacking ensemble of SVM + KNN + RF

Emotion	Acc	W-Prec	W-Rec	W-F1	M-Prec	M-Rec	M-F1
Angry	0.95	0.95	0.9505	0.95	0.95	0.9505	0.95
Happy	0.9571	0.9571	0.958	0.9571	0.9571	0.958	0.9571
Neutral	0.9657	0.9658	0.9663	0.9657	0.9658	0.9663	0.9657
Sad	0.9714	0.9713	0.9718	0.9714	0.9713	0.9718	0.9714
Surprise	0.9643	0.9643	0.9646	0.9643	0.9643	0.9646	0.9643
Average	0.9617	0.9617	0.9622	0.9617	0.9617	0.9622	0.9617

Table 5. Classifier-wise Accuracy Across Emotions

Avg = Average accuracy across emotions

Model	Angry	Happy	Neutral	Sad	Surprise	Avg
SVM	0.9486	0.9636	0.9657	0.9707	0.9579	0.9613
KNN	0.9343	0.9443	0.9486	0.9679	0.9436	0.9477
RF	0.9193	0.925	0.9336	0.9421	0.9243	0.9289
Proposed Stacking	0.95	0.9571	0.9657	0.9714	0.9643	0.9617

Statistical Evaluation

Figure 2 presents the average Friedman ranks computed for the individual classifiers and the proposed stacking framework as per accuracy. Among the base models, RF achieves the lowest rank (6), indicating

the most consistent performance across emotional conditions. KNN follows with a moderate rank of 12, while SVM records a higher rank of 19.5, reflecting greater variability in its performance. In contrast, the proposed stacking approach attains the highest rank value of 22.5, clearly demonstrating its superior overall effectiveness. The substantial improvement observed in the ensemble model highlights the advantage of combining complementary decision patterns from multiple classifiers, resulting in a more robust and reliable speaker identification system.

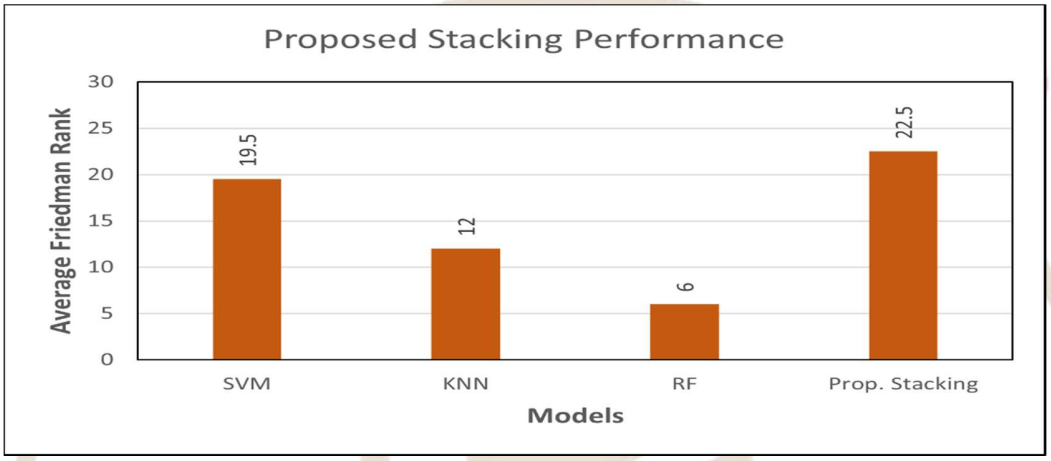


Figure 2: Friedman Ranking Comparison on Existing Classifier

Figure 3 illustrates the average Friedman ranks obtained across different emotional categories, providing insight into how emotion influences model performance. The Angry emotion achieves the lowest rank value of 4, indicating the most consistent and stable recognition performance among all categories. This is followed by Surprise with a rank of 10 and Happy with 12, reflecting moderate variability. In contrast, Neutral and Sad exhibit higher ranks of 20 and 24 respectively, showing that these emotions pose greater challenges for speaker discrimination. The overall average rank of 14 summarizes the general performance trend across emotions. These results highlight that emotional states significantly impact model behavior, with certain emotions such as Angry and Surprise being more easily identifiable, while Neutral and Sad introduce more variability and complexity to the recognition task.

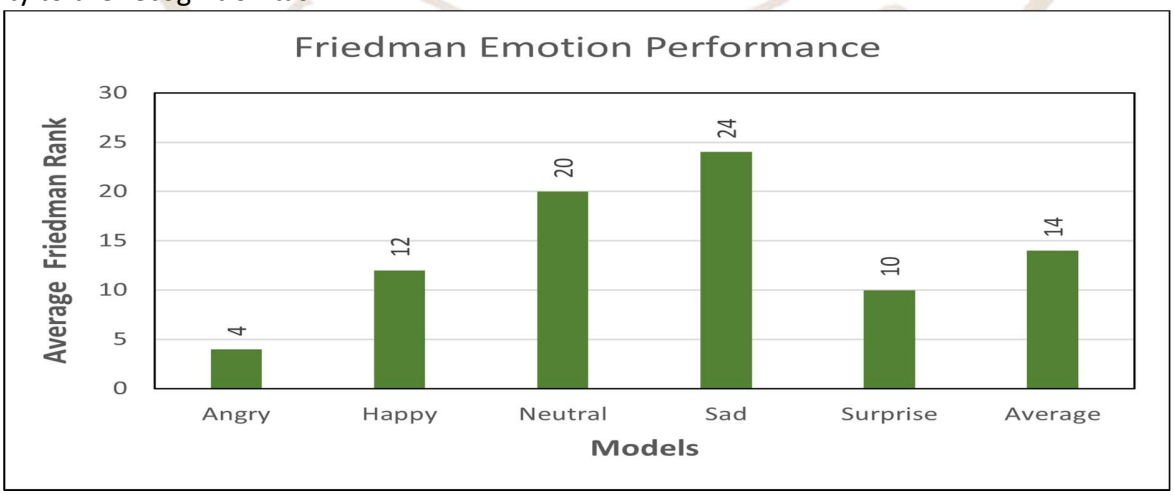


Figure 3: Friedman Ranking Comparison on Emotions

AD EDUXIAN JOURNAL
A Quarterly Multidisciplinary Blind Peer Reviewed & Refereed Online International Journal
Google Scholar Indexing PIF:1.048(I2OR) & 3.125 (IIFS)
Vol (2), Issue (3), Oct-Dec 2025 ISSN: 3048-7951

Additional Result Evaluation: Figure 5 of confusion matrices illustrates the performance of the stacking ensemble across five emotional categories: Angry, Happy, Neutral, Sad, and Surprise. In all matrices, the strong diagonal dominance indicates highly accurate identification of speakers, with mis- classifications remaining minimal and sparsely distributed across non-diagonal cells. The Angry and Surprise emotions show the highest clarity, with dense diagonal blocks and almost negligible confusion between speakers, reflecting the system's strong discrimina- tive ability under these conditions. Happy and Neutral emotions exhibit slightly more scattered off-diagonal entries, suggesting increased variability but still maintaining strong overall recognition accuracy. The Sad emotion shows the greatest degree of dispersion, indicating that this emotional state introduces more overlap among speaker characteris- tics. Overall, the confusion matrices demonstrate that the proposed stacking framework maintains consistently high speaker identification performance across emotional condi- tions, with only minor variability in difficulty among emotions.

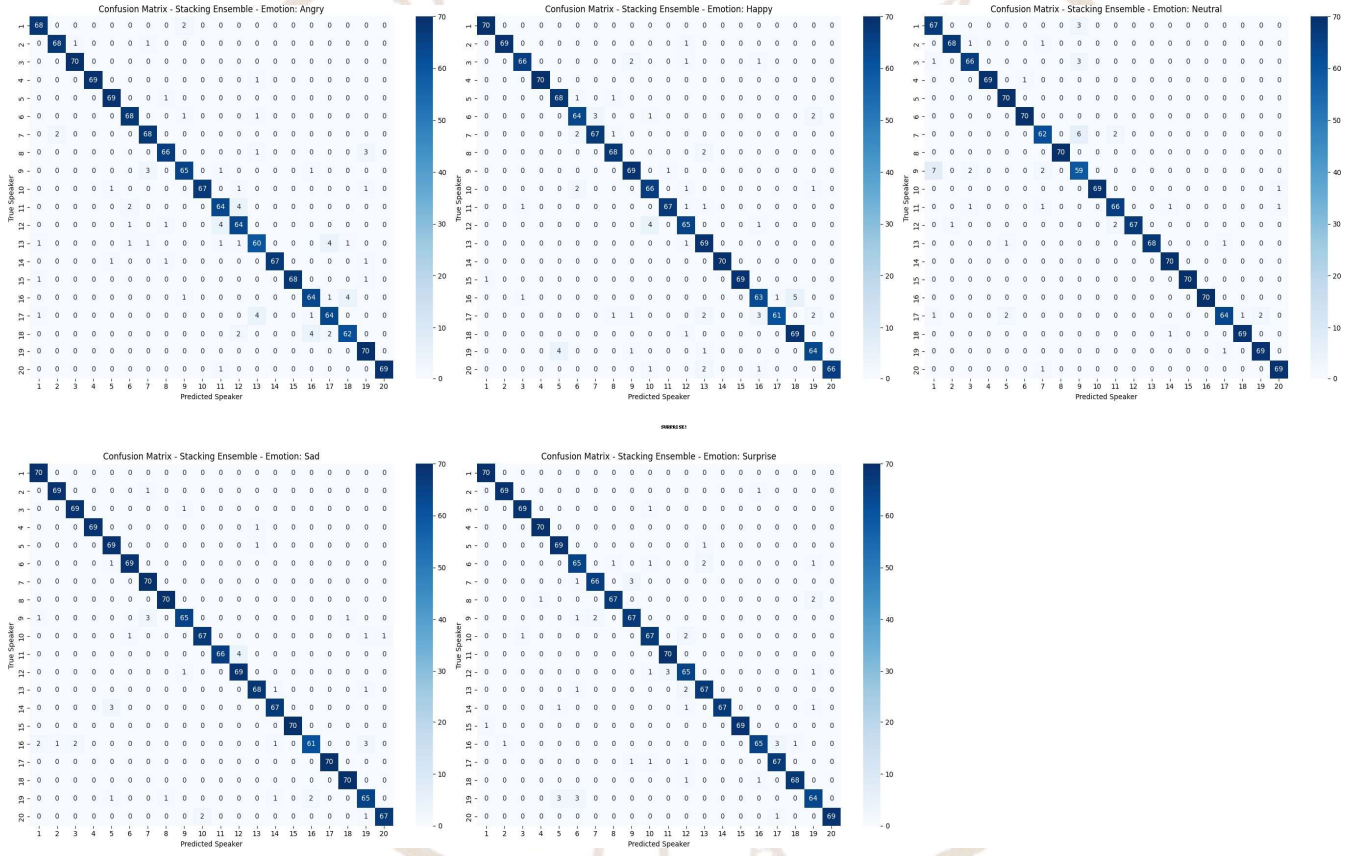


Figure 4: Confusion Matrix for Speaker Identification

Performance Comparison with SOTA Approaches: We evaluate the effectiveness of the proposed model by benchmarking it against several state-of-the-art (SOTA) speaker identification techniques, as summarized in Table 6. The comparison highlights the accuracies reported by leading existing approaches on the ESD emotional speech dataset, where Speech VGG by Hamsa et al. (2022), a Feedforward Neural Network by Oo et al. (2024), and the Caps Net-M model by Nassif et al. (2022) achieved decent accuracies. In contrast, the proposed stacking-based ensemble model in- targeting SVM, KNN, and RF classifiers achieves a superior accuracy of 96.17%, demonstrating its enhanced capability to distinguish speakers even under emotional variability. This substantial improvement clearly underscores the robustness, stability, and overall superior performance of the proposed system compared to existing SOTA methods.

The bar chart 5 visually compares the performance of several speaker identification methods evaluated on an emotional speech dataset. Each bar represents a distinct approach, making it easy to observe the relative differences in their effectiveness. The proposed model stands out prominently, with its bar rising significantly higher than the rest, indicating a clear improvement over the existing techniques. This visual trend highlights the enhanced robustness and reliability of the proposed system in handling emotionally varied speech, demonstrating its ability to outperform prior methods and deliver more accurate speaker identification results.

Table 6. Accuracy Comparison with Existing SOTA Methods on ESD Dataset

SOTA = State-of-the-Art; Acc = Accuracy (%)

Reference	Model	Accuracy (%)
Hamsa et al. (2022)	Speech VGG	86.67
Oo et al. (2024)	Feedforward Neural Network	89.39
Nassif et al. (2022)	CapsNet-M	89.89
Proposed Work	Ensemble (SVM + KNN + RF)	96.17

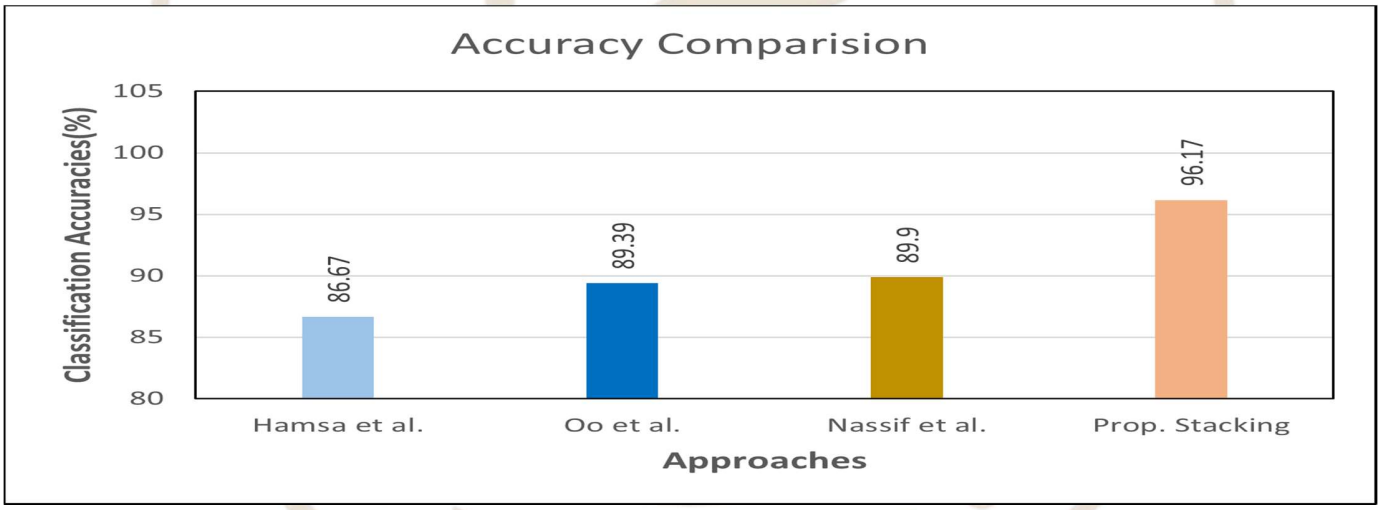


Figure 5: Accuracy Comparison of the Proposed Method with Existing

Conclusion: The evaluation on the ESD dataset demonstrates that the proposed stacking-based ensem- ble architecture is highly effective for speaker identification in emotional environments. By integrating SVM, KNN, and RF classifiers, the model achieves a strong accuracy of 96.17%, highlighting its ability to capture emotion-invariant speaker characteristics and deliver stable performance. Although the proposed system shows promising results, there remains significant scope for further enhancement. Future work may explore ad- vanced fusion and hybrid fusion strategies to combine complementary acoustic cues more effectively. Additionally, hybrid ensemble techniques can be investigated to improve ro- bustness under complex emotional variations.

References:

1. Sandra Pruzansky. Pattern-matching procedure for automatic talker recognition. The Journal of the Acoustical Society of America, 35(3):354–358, 1963.
2. Sanghamitra Mohanty and Basanta Kumar Swain. Speaker identification using svm during oriya speech recognition. International Journal of Image, Graphics and Signal Processing, 7(10):28, 2015.

3. Sreenivas Sremath Tirumala, Seyed Reza Shahamiri, Abhimanyu Singh Garhwal, and Ruili Wang. Speaker identification features extraction methods: A systematic review. Expert Systems with Applications, 90:250–271, 2017.
4. Khine Zin Oo, Lwin Nyein Thu, and Zaw Htet Aung. Enhancement of speaker iden- tification system based on voice active detection techniques using machine learning. In 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon), pages 889–893. IEEE, 2024.
5. Vincent Wan and William M Campbell. Support vector machines for speaker verifi- cation and identification. In Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No. 00TH8501), volume 2, pages 775–784. IEEE, 2000.
6. Shai Fine, Jiri Navratil, and Ramesh A Gopinath. A hybrid gmm/svm approach to speaker identification. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), volume 1, pages 417–420. IEEE, 2001.
7. Pedro J Moreno and Purdy Ho. A new svm approach to speaker identification and verification using probabilistic distance kernels. In INTERSPEECH, pages 2965– 2968, 2003.
8. Vijendra Raj Apsingekar and Phillip L De Leon. Support vector machine based speaker identification systems using gmm parameters. In 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, pages 1766–1769. IEEE, 2009.
9. Douglas A Reynolds. Experimental evaluation of features for robust speaker identi- fication. IEEE Transactions on Speech and Audio Processing, 2(4):639–643, 2002.
10. Ravi P Ramachandran, Kevin R Farrell, Roopashri Ramachandran, and Richard J Mammone. Speaker recognition—general classifier approaches and data fusion meth- ods. Pattern recognition, 35(12):2801–2821, 2002.
11. Noor Almaadeed, Amar Aggoun, and Abbes Amira. Speaker identification using multimodal neural networks and wavelet analysis. Iet Biometrics, 4(1):18–28, 2015.
12. V Karthikeyan and S Suja Priyadharsini. A strong hybrid adaboost classification algorithm for speaker recognition. S¯adhan¯a, 46(3):138, 2021.
13. Xingmei Wang, Jiaxiang Meng, Bin Wen, and Fuzhao Xue. Racp: A network with attention corrected prototype for few-shot speaker recognition using indefinite dis- tance metric. Neurocomputing, 490:283–294, 2022.
14. Rasha H Ali, Mohammed Najm Abdullah, and Buthainah F Abed. The identification and localization of speaker using fusion techniques and machine learning techniques. Evolutionary Intelligence, 17(1):133–149, 2024.
15. Mahesh K Singh. Feature extraction and classification efficiency analysis using machine learning approach for speech signal. Multimedia Tools and Applications, 83(16):47069–47084, 2024.
16. Anett Antony and R Gopikakumari. Speaker identification based on combination of mfcc and umrt based features. Procedia computer science, 143:250–257, 2018.
17. Nguyen Nang An, Nguyen Quang Thanh, and Yanbing Liu. Deep cnns with self- attention for speaker identification. IEEE access, 7:85327–85337, 2019.
18. Rashid Jahangir, Ying Wah Teh, Henry Friday Nweke, Ghulam Mujtaba, Mo- hammed Ali Al-Garadi, and Ihsan Ali. Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. Expert Systems with Applications, 171:114591, 2021.
19. Shibani Hamsa, Ismail Shahin, Youssef Iraqi, Ernesto Damiani, and Naoufel Werghi. Speaker identification from emotional and noisy speech data using learned voice segregation and speech vgg. arXiv preprint arXiv:2210.12701, 2022.
20. Ali Bou Nassif, Ismail Shahin, Ashraf Elnagar, Divya Velayudhan, Adi Alhudhaif, and Kemal Polat. Emotional speaker identification using a novel capsule nets model. Expert Systems with Applications, 193:116469, 2022.

21. Nirupam Shome, Banala Saritha, Richik Kashyap, and Rabul Hussain Laskar. A ro- bust dnn model for text-independent speaker identification using non-speaker embed- dings in diverse data conditions. Neural Computing and Applications, 35(26):18933– 18947, 2023.
22. Dongdong Li, Zhuo Yang, Jinlin Liu, Hai Yang, and Zhe Wang. Emotion embedding framework with emotional self-attention mechanism for speaker recognition. Expert Systems with Applications, 238:122244, 2024.
23. Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional voice conversion: Theory, databases and esd. Speech Communication, 137:1–18, 2022.
24. Shibani Hamsa, Ismail Shahin, Youssef Iraqi, Ernesto Damiani, Ali Bou Nassif, and Naoufel Werghi. Speaker identification from emotional and noisy speech using learned voice segregation and speech vgg. Preprint on SSRN, 2022.