International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(22s) 2026 pp. 1414-1432
https://doi.org/10.70917/ijcisim-2026-5539
Article

MARGIN–RELIABILITY GATED ENSEMBLE FOR ROBUST SPEECH ACCENT RECOGNITION USING HANDCRAFTED ACOUSTIC FEATURES

Prabagaran D[1], Raja B[2], Vinothkumar A[3], Anand M[4]

[1]Research Scholar, 2,3,4Professor
[1,2,3,4]Department of Electronics and Communication Engineering, Dr. M. G. R. Educational and Research Institute, Chennai.
[1]praba.arni@gmail.com, [2]raja.ece@drmgrdu.ac.in, [3]vinothkumar.ece@drmgrdu.ac.in, [4]anand.ece@drmgrdu.ac.in

Abstract: Speech accent recognition is an emerging area of speech technology with applications in human–computer interaction, automated transcription, and language learning systems. Although deep learning has recently dominated the field, conventional machine learning models based on handcrafted acoustic features remain highly relevant due to their interpretability, lower computational cost, and suitability in resource-constrained environments. This paper proposes the Margin–Reliability Gated Ensemble (MR-GE), a novel machine learning framework that integrates calibrated margins from multiple baseline classifiers (Support Vector Machines, Random Forests, k-Nearest Neighbors, and Gaussian Mixture Models), accent-aware reliability weights derived from per-class validation profiles, and sparse non-negative feature-group gating across MFCC, PLP, prosodic, and spectral descriptors. The framework is systematically evaluated against baseline classifiers and optimized through dimensionality reduction and feature selection methods such as PCA, LDA, mRMR, and ReliefF. Experimental results on multi-accent speech datasets demonstrate that MR-GE achieves significant improvements in recognition accuracy, F1-score, and calibration reliability compared to individual classifiers. These findings establish a robust baseline-to-optimized machine learning pipeline for accent recognition and provide a foundation for future integration with deep learning approaches.
Keywords: Speech Accent Recognition; Machine Learning; Handcrafted Acoustic Features; Feature Selection; Dimensionality Reduction; Ensemble Learning; Margin–Reliability Gated Ensemble

1. INTRODUCTION

Speech accent recognition is an important subtask in automatic speech processing, where the goal is to identify the accent of a speaker based on short audio samples. Accurate accent recognition plays a critical role in enhancing the robustness of speech recognition systems, improving speaker adaptation, and enabling personalized language learning applications [1]. However, accent variability is influenced by several linguistic and phonetic factors, making it a challenging problem [2].

Although deep neural networks and attention-based architectures dominate current speech research, they often require large-scale annotated datasets and high computational resources [3]. In contrast, conventional machine learning models with handcrafted acoustic features continue to provide strong baselines, especially for resource-limited environments and underexplored datasets [4]. Handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP), prosodic measures, and spectral descriptors have historically been effective for capturing accent-specific patterns, yet their performance is strongly influenced by the choice of classifiers and feature optimization methods [5].

Copyright: © 2026 by the authors

This paper introduces the Margin–Reliability Gated Ensemble (MR-GE), a novel machine learning framework designed to enhance accent recognition accuracy while maintaining computational efficiency [6]. The framework builds on baseline classifiers by introducing margin calibration, accent-aware reliability weighting, and sparse feature-group gating [7]. By systematically integrating multiple conventional classifiers [8] and feature optimization techniques [9], MR-GE establishes a new optimized benchmark for accent recognition that bridges the gap between traditional ML baselines and more complex deep learning models [10]. The main contributions of this work are as follows:

- Systematic Baseline Establishment: Development of baseline machine learning models for accent recognition using handcrafted acoustic features, including MFCC, PLP, prosodic, and spectral descriptors.
- Novel Ensemble Algorithm (MR-GE): Introduction of the Margin–Reliability Gated Ensemble, which combines margin calibration, accent-aware reliability weights, and sparse non-negative feature-group gating for robust classification.
- Feature Optimization Framework: Comparative evaluation of multiple feature selection and dimensionality reduction techniques (PCA, LDA, mRMR, ReliefF) to highlight their impact on accent recognition performance.

The remainder of this paper is organized as follows. Section 2 reviews related works on speech accent recognition, focusing on conventional machine learning models, handcrafted acoustic features, feature optimization strategies, and ensemble-based approaches. Section 3 presents the proposed framework in detail, including preprocessing, feature extraction, feature selection and dimensionality reduction, baseline classifiers, and the design of the Margin–Reliability Gated Ensemble (MR-GE). Section 4 describes the experimental setup, covering dataset characteristics, preprocessing parameters, training protocols and experimental results. Finally, Section 5 concludes the paper by summarizing key contributions and outlining directions for future research, particularly the integration of deep learning architectures with the proposed ensemble framework.

2. RELATED WORKS

Accent classification research has focused on improving ASR by analyzing datasets, applying preprocessing, extracting acoustic features, and testing classification models across traditional ML and DL approaches. Comparative evaluations highlight the strengths of k-NN, CNNs, ResNet variants, and transformer-based architectures in handling diverse accents [11]. Speech emotion recognition has been approached through MFCC-based feature extraction and spectrogram analysis, applying both machine learning and deep learning models. Ensemble classifiers, particularly SVM combined with boosting and voting methods, have been shown to enhance recognition performance [12].

Automated speech disorder detection integrates mel-spectrogram feature generation with deep feature extractors such as ResNet to enhance diagnostic precision. Ensemble learning, combining CatBoost, XGBoost, and SVM, is applied to strengthen prediction reliability [13]. Voice-based alcohol detection leverages acoustic features to differentiate drinkers from non-drinkers using ensemble learning strategies. Models including Naive Bayes, k-NN, decision trees, and SVM are stacked to boost classification capability [14].

Accent presence identification employs a combination of prosodic and spectral features extracted from speech signals. Multiple machine learning classifiers such as KNN, SVM, decision trees, and logistic regression are evaluated for their ability to detect accents [15]. Accent-based ASR for Nigerian English utilizes MFCCs for preprocessing and a hybrid RBM-LSTM framework for sequence modeling. The system processes speech from multiple ethnic groups, with training guided by backpropagation through time [16].

Accent classification across major English varieties applies CNN, SVM, and random forest models using the Mozilla Common Voice dataset. Performance analysis reveals SVM's superiority in differentiating subtle regional variations compared to deep models [17]. A transfer learning framework enhances speech emotion recognition by integrating accent identification knowledge. DNNs trained on accent features are coupled with convolutional recurrent

autoencoders for feature mapping [18]. Age and gender recognition from speech applies MFCC features with multiple base classifiers including KNN, SVM, RF, LR, and decision trees. Ensemble strategies such as stacking and majority voting are used to maximize accuracy [19].

Speaker background profiling investigates accent classification through datasets, extracted features, and comparative analyses of classifiers. A systematic review highlights the methods and performance measures used in recent works [20]. Tonal contrast recognition in the Chokri language applies supervised ML algorithms and ANN to acoustic features like f0 height and direction. The models are evaluated for their ability to capture subtle tonal variations across genders [21].

ASR performance is enhanced using ensemble strategies with large language models through a novel Priority-Encoder Ensemble (PE-Ensemble). This approach employs a meta-learning decider to dynamically select optimal models during inference [22]. Accent detection for demographic profiling uses the Speech Accent Archive dataset to classify speakers' linguistic backgrounds. Recordings of scripted English phrases are processed to identify language-based foreign accent markers [23].

Accent identification between native French and American speakers is approached using phoneme-based classification. MFCC features with SVM are compared to spectrogram-based ResNet classifiers, highlighting the superiority of deep visual representations [24]. Accents-based speech recognition for e-learning applies hybrid RBM and LSTM neural architectures to address comprehension barriers. This integration exploits acoustic and temporal modeling to improve recognition across diverse learner groups [25].

3. PROPOSED MODEL

The proposed model introduces the Margin–Reliability Gated Ensemble (MR-GE) framework for robust speech accent recognition using handcrafted acoustic features as shown in Fig 1. The pipeline begins with preprocessing of raw speech signals through DC removal, pre-emphasis, framing, windowing, and normalization, followed by the extraction of MFCC, PLP, prosodic, and spectral descriptors. To enhance discriminability, feature optimization is performed using filter-based methods such as mRMR and ReliefF, wrapper approaches like Sequential Forward Selection, and dimensionality reduction techniques including PCA and LDA.

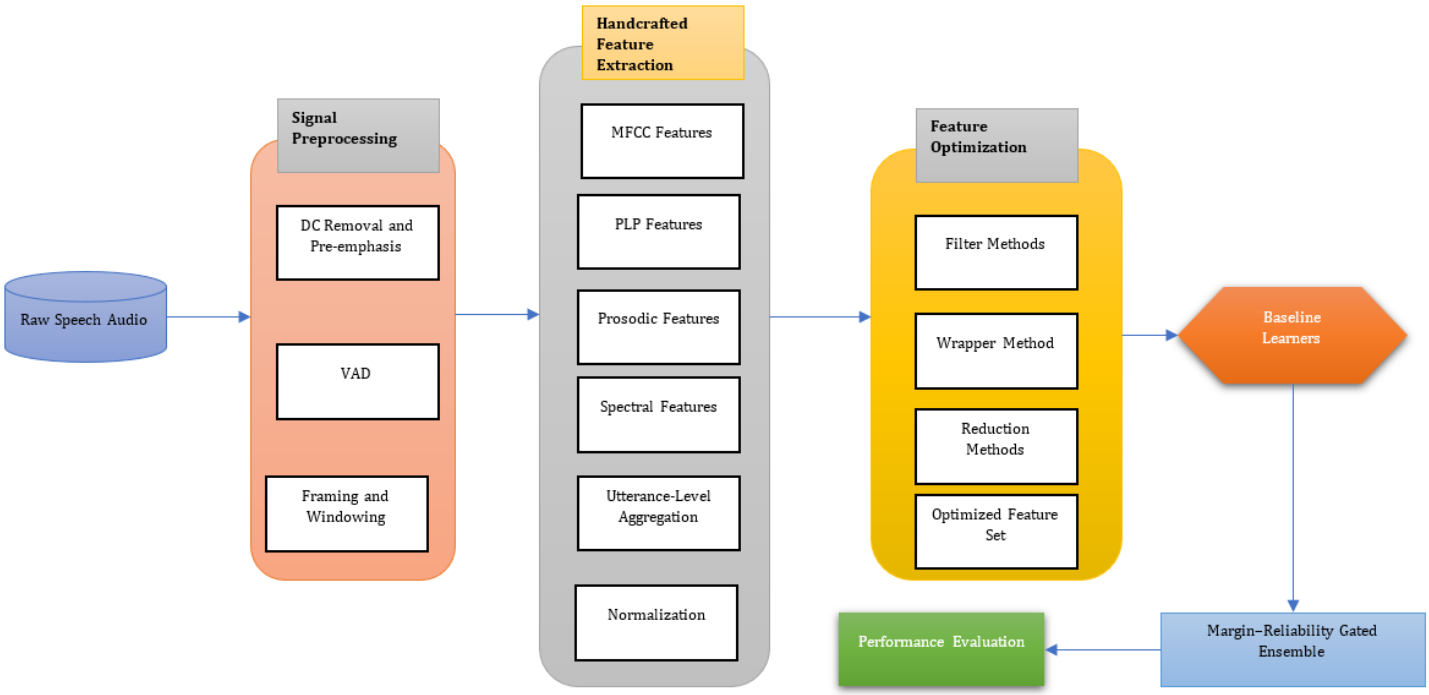


Figure 1: Overall Architecture of Proposed Margin–Reliability Gated Ensemble Model

Conventional classifiers, namely Support Vector Machines, Random Forests, k-Nearest Neighbors, and Gaussian Mixture Models, are trained on optimized feature groups and their outputs are calibrated into probabilistic scores. MR-GE then fuses these outputs by assigning accent-aware reliability weights derived from class-specific validation performance, global weights to balance individual classifiers, and sparse non-negative gates to highlight the most informative feature groups. The final decision is obtained through a reliability- and gate-weighted consensus score, ensuring interpretability, robustness against accent overlap, and improved recognition accuracy compared to individual classifiers or unoptimized baselines.

Step 1: Data Acquisition & Partitioning

This step focuses on preparing a balanced and standardized dataset for speech accent recognition. Multi-accent corpora are curated such that each accent has at least 5–10 hours of recordings to ensure sufficient coverage. All audio is resampled to a uniform format of 16 kHz, 16-bit PCM, mono, making signals directly comparable. The dataset is then split into 80% training and 20% testing sets in a speaker-independent and stratified manner to avoid bias and leakage.

Dataset Description

In this study, three publicly available corpora were employed to ensure diversity, reproducibility, and robustness of the proposed model. The **Mozilla Common Voice (CV)** dataset provides one of the largest multilingual and multi-accent collections, with thousands of speakers contributing across diverse English accents such as Indian, American, British, Australian, and African. The **VoxForge corpus** was selected as it offers crowd-sourced accented English recordings, which are widely used for benchmarking accent recognition systems. To further strengthen generalization, the **Speech Accent Archive (SAA)** was incorporated, containing carefully curated samples where speakers from more than 100 first-language backgrounds read the same English passage, making accent differences highly perceptible. Together, these datasets offer both large-scale, natural speech (CV), balanced yet smaller-scale accented speech (VoxForge), and controlled accent variability (SAA), providing a comprehensive benchmark for evaluating baseline and optimized models.

Table 1: Summary of Speech Accent Datasets

Dataset Name	Accents Covered	No. of Speakers	Total Hours	Unique Features
Mozilla Common Voice (English subset)	US, UK, Indian, Australian, African, Canadian, Others	> 7,000	~1,800 hrs	Large-scale, diverse, open-source, ongoing updates
VoxForge	US English, UK English, Canadian, Australian, Indian	~1,200	~100 hrs	Crowd-sourced, widely used in accent recognition baselines
Speech Accent Archive	>100 first-language backgrounds (e.g., Chinese, Hindi, Arabic, Russian, Spanish)	~2,000	~80 hrs	All speakers read the same paragraph, making accents directly comparable

Table 1 presents the key characteristics of the datasets used in this study, including accents covered, number of speakers, total duration, and unique features, highlighting their suitability for robust accent recognition research.

We denote the set of accents as $C = \{1, 2, , \dots, C\}$ with utterances u_c for each accent. The total hours for accent c is

$$T_c = \sum_{u \in u_c} \frac{T_u}{3600}, T_c \geq T_{min} \quad (1)$$

Audio is resampled to 16 kHz and quantized to 16-bit PCM, while ensuring stratified splits with

$$\frac{T_c^{train}}{T_c} \approx 0.7, \frac{T_c^{oval}}{T_c} \approx 0.7, \frac{T_c^{test}}{T_c} \approx 0.2 \quad (2)$$

Step 2: Signal Preprocessing

The raw speech signals are first cleaned by removing DC offset and applying pre-emphasis to enhance higher frequencies. Voice Activity Detection (VAD) is then performed using energy and zero-crossing measures to discard silent or non-speech regions. Finally, the signal is framed into short segments with Hamming windowing and normalized to unit amplitude for consistent analysis.

1. DC Removal and Pre-emphasis

To eliminate DC offset, each signal $x[n]$ is mean-normalized:

$$x'[n] = x[n] - \frac{1}{N} \sum_{k=0}^{N-1} x[k] \quad (3)$$

Then a pre-emphasis filter with $\alpha = 0.97$ is applied:

$$x''[n] = x'[n] - \alpha x'[n-1] \quad (4)$$

2. Voice Activity Detection (VAD)

Speech frames are detected using short-time energy and zero-crossing rate. The energy of a frame f is:

$$E_f = \sum_{n=1}^{N} (x''[n])^2, \qquad ZCR_f = \frac{1}{2} \sum_{k=0}^{N-1} |sgn(x''[n]) - sgn(x''[n-1])| \quad (5)$$

Frames with $E_f < \theta_E$ and $ZCR_f < \theta_Z$ are discarded as silence or noise.

3. Framing and Windowing

The signal is divided into overlapping frames of 25 ms with a 10 ms hop. Each frame f is multiplied by a Hamming window:

$$w[n] = 0.54 - 0.46 \cos\cos\left(\frac{2\pi n}{N-1}\right), x_f[n] = x''[n] \cdot w[n] \quad (6)$$

This reduces spectral leakage in the Fourier domain.

4. Amplitude Normalization

Each utterance is normalized of unit root mean square (RMS) energy to ensure scale invariance:

$$x_{norm}[n] = \frac{x_f[n]}{\sqrt{\frac{1}{N}\sum_{k=0}^{N-1} \left(x_f[k]\right)^2}} \quad (7)$$

This guarantees consistent loudness across all recordings.

Step 3: Handcrafted Feature Extraction (Feature Groups)

From each preprocessed frame, complementary acoustic features are computed and then aggregated at the utterance level using statistics such as mean, standard deviation, and percentiles. The final utterance vector is standardized with z-score normalization to ensure comparability across speakers and accents.

1. MFCC Features(G_1)

Mel-Frequency Cepstral Coefficients (MFCCs) are derived from the log-mel spectrum. If $X[k]$ is the magnitude spectrum and $M_i[k]$ is the mel filter bank, then:

$$E_i = \sum_k |X[k]^2| M_i[k], \qquad c_j = \sum_{i=1}^{I} log\ log\ (E_i)\ Cos\left(\frac{\pi j}{I}(i-0.5)\right) \tag{8}$$

With 13 cepstra plus first and second, we obtain 39 dimensions per frame.

2. PLP Features(G_2)

Perceptual Linear Prediction (PLP) approximates the auditory spectrum using all-pole model. The LPC coefficients a_p are estimated by minimizing prediction error:

$$\epsilon = \sum_n \left((x) + \sum_{p=1}^{P} a_p x[n-p]\right)^2 \tag{9}$$

Typically, 13 PLP coefficients with $\Delta, \Delta\Delta$ are used, optionally with RASTA filtering for channel normalization.

3. Prosodic Features(G_3)

Prosodic cues are derived from pitch, energy and temporal measures. Pitch is estimated via autocorrelation:

$$R(T) = \sum_n x[n]x[n-T], \qquad F_0 = \frac{f_s}{arg\ arg\ R(T)} \tag{10}$$

Statistical descriptors include mean, standard deviation and range of F_0, energy statistics, jitter, shimmer, and voiced/unvoiced ratios.

4. Spectral Features(G_4)

Spectral descriptors capture shape and spread of the spectrum. Examples include:

- Spectral centroid:

$$C = \frac{\sum_k k|X[k]|}{\sum_k |X[k]|} \tag{11}$$

- Spectral flux:

$$F = \sum_k \left(|X_t[k]| - \sum_k |X_{t-1}[k]|\right)^2 \tag{12}$$

- Roll-off: the frequency below which 85% or 95% of energy lies. Formant Frequencies F1, F2, F3 are extracted using LPC roots.

5. Utterance-Level Aggregation

For each feature dimension d, statistics are computed:

$$\mu_d = \frac{1}{N}\sum_{n=1}^{N} f_d[n], \quad \sigma_d = \sqrt{\frac{1}{N}\sum_{n=1}^{N} (f_d[n] - \mu_d)^2} \tag{13}$$

Percentiles are also recorded, yielding a fixed utterance-level vector.

6. Normalization

Finally, z-score normalization is applied per feature dimension using training statistics:

$$z_d = \frac{f_d - \mu_d^{train}}{\sigma_d^{train}} \tag{14}$$

$Algorithm\ 0: End-to-End\ Feature\ Pipeline\ (Context)$
$Input: Raw\ utterance\ waveform\ x_raw, sampling\ rate\ fs_in$
$Output: Utterance-level\ normalized\ vector\ z$
$1: x_std \leftarrow StandardizeAudio(x_raw, fs_in \rightarrow 16\ kHz, mono, 16-bit)$
$2: x_pre \leftarrow Preprocess(x_std) \quad \#\ DC\ removal, pre-emphasis, VAD$
$3: frames \leftarrow FrameAndWindow(x_pre, 25\ ms, 10\ ms, Hamming)$
$4: z \leftarrow HandcraftedFeatureExtractionAndAggregation(frames, fs = 16\ kHz)\ \#\ Algorithm\ 1$
$5: return\ z$

Step 4: Feature Optimization (Selection/Reduction)

After feature extraction, the dimensionality of the utterance-level vector can be very high, leading to redundancy and computational overhead. To address this, we apply feature optimization through filter, wrapper, and reduction methods, retaining the configuration that yields the best macro-F1 score on the validation set.

1. Filter Methods

(a) mRMR (Minimum Redundancy Maximum Relevance)

For feature f_j, relevance with respect to class labels c is measured by mutual information:

$$I(f_j; c) = \sum_{f_j} \sum_{c} P(f_j, c)\ log\ log\ \frac{P(f_j, c)}{P(f_j)P(c)} \tag{15}$$

Redundancy between features f_j, f_k:

$$(f_j, f_k) = I(f_j;\ f_k) \tag{16}$$

mRMR selects features maximizing

$$\left[\frac{1}{|S|}\sum_{f_j \in S} \quad I(f_j; c) - \frac{1}{|S|^2}\sum_{f_j, f_k \in S} \quad I(f_j; f_k)\right] \tag{17}$$

Where S is the subset of features.

(b) ReliefF

Weights are updated by comparing nearest-hit and nearest-miss samples. For feature F_j:

$$W(f_j) \leftarrow W(f_j) - \frac{1}{m}\sum_{i=1}^{m} \quad diff\left(f_j, x_i, hit(x_i)\right) + \frac{1}{m}\sum_{i=1}^{m} \quad diff\left(f_j, x_i, miss(x_i)\right) \tag{18}$$

Features with the highest $W(f_j)$ are selected.

2. Wrapper Method

Sequential Forward Selection (SFS)

Start with empty set $S = \emptyset$. Iteratively add feature f^* that maximizes validation accuracy:

$$f^* = arg\ arg\ Acc(S \cup \{f_j\}) \tag{19}$$

Until the budget constraint $|S| \leq 50 is\ reached$.

3. Reduction Methods

(a) Principal Component Analysis (PCA)

Transform Features into uncorrelated components. Covariance matrix:

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N} \quad (x_i - \mu)(x_i - \mu)^T \tag{20}$$

Eigen decomposition: $\sum \qquad vj = \lambda_j v_j$.

Retain top k components such that

$$\frac{\sum_{j=1}^{k} \quad \lambda_j}{\sum_{j=1}^{d} \quad \lambda_j} \geq \theta, \theta \in \{0.95, 0.95\} \tag{21}$$

(b) Linear Discriminant Analysis (LDA)

Compute between-class scatter S_B and within-class scatter S_W:

$$S_B = \sum_{c=1}^{C} \quad N_c(\mu_C - \mu)(\mu_C - \mu)^T, S_W = \sum_{c=1}^{C} \quad \sum_{x_i \in c} \quad (x_i - \mu_c)(x_i - \mu_c)^T \tag{22}$$

Projection matrix W maximizes:

$$W^* = arg\ arg\ \frac{|W^T S_B W|}{|W^T S_W W|} \tag{23}$$

Yielding at most C-1 compoents.

4. Optimized Feature Set

For each feature group, the chosen method (mRMR, ReliefF, SFS, PCA, LDA) and dimensionality k are those producing the highest **macro-F1** on the validation split. An **All-Features track** (no selection) is also retained as a baseline for comparison.

Table 2: Feature Optimization Techniques and Mathematical Formulations

Category	**Method**	**Objective / Criterion**	**Dimensionality Rule**
Filter	mRMR	Maximize relevance, minimize redundancy	S
	ReliefF	Update feature weights by nearest hit/miss comparison	Rank & choose top-kk
Wrapper	SFS	Incrementally add features maximizing classifier accuracy	Budget ≤ 50 features
Reduction	PCA	Retain components with highest variance	Choose top-kk PCs
	LDA	Maximize between-class separability	W^T S_B W

This table summarizes the feature optimization methods applied to reduce redundancy and improve accent discrimination. It highlights filter-based approaches (mRMR, ReliefF), a wrapper method (SFS), and reduction techniques (PCA, LDA). Each method is described with its optimization objective, key mathematical formulation, and rule for selecting dimensions. Together, these strategies ensure compact yet highly discriminative feature sets for robust classification.

Algorithm 2: Feature Optimization
Input: Feature groups {G1, G2, G3, G4}, Labels Y
Output: Optimized feature sets {G1*, G2*, G3*, G4*}
1: For each group Gi in {G1, G2, G3, G4} do
2: # Apply Filter methods
3: Rank features using mRMR and ReliefF
4: Select top−k features (k ∈ {32, 64, 128, 256})
5:
6: # Apply Wrapper method
7: Perform Sequential Forward Selection (SFS)
8: Retain subset with ≤ 50 features
9:
10: # Apply Reduction methods
11: Perform PCA, keep components with ≥95% variance
12: Perform LDA, keep at most C−1 components
13:
14: # Validation selection

15: Evaluate each method using macro-F1 on validation set
16: Choose best-performing subset as Gi*
17: End For
18:
19: Return optimized feature sets {G1*, G2*, G3*, G4*}

Step 5: Baseline Learners

"To establish reference benchmarks, conventional classifiers including SVM (RBF kernel), Random Forests, kNN, and Gaussian Mixture Models were trained on individual feature groups (G1–G4) and concatenated vectors. Hyperparameters followed standard grids (e.g., C ∈ {1,10,100} for SVM; 200–500 trees for RF). Classifier outputs were calibrated via Platt scaling, temperature scaling, or isotonic regression to yield comparable probability estimates. These baselines provide a reference for assessing the improvements introduced by the proposed MR–GE model."

Algorithm 3: Train and Calibrate Baselines Per Group
Input: Feature groups {G1..G4, Gall}, labels Y, validation split Val
Output: Calibrated models M = {SVM, RF, kNN, GMM} per group
1: for each group G in {G1..G4, Gall} do
2: # Fit models with given grids
3: $svm_G \leftarrow TrainSVM_RBF(G;\ C \in \{1,10,100\}, \gamma \in \{1/d, 1/(2d)\})$
4: $rf_G \leftarrow TrainRF(G;\ trees \in \{200,500\}, max_depth \in \{None, 20\})$
5: $knn_G \leftarrow TrainKNN(G;\ k \in \{3,5,7,11\}, metric = Euclidean)$
6: $gmm_G \leftarrow TrainGMM(G;\ diag, components \in \{4,8\}\ per\ class)$
7:
8: $\#\ Convert\ scores\ to\ calibrated\ probabilities\ on\ Val$
9: $svm_G.prob \leftarrow PlattCalibrate(svm_G, Val)$
10: $rf_G.prob \leftarrow TempScale(rf_G, Val)$
11: $knn_G.prob \leftarrow IsotonicCalibrate(DistWeightedProbs(knn_G), Val)$
12: $gmm_G.prob \leftarrow PosteriorProbs(gmm_G);\ gmm_G.prob \leftarrow TempScale(gmm_G.prob, Val)$
13:

14: $Save\ \{svm_G, rf_G, knn_G, gmm_G\}\ to\ M$
15: end for
16: return M

Step 6: Proposed Algorithm — MR-GE (Margin–Reliability Gated Ensemble)

The proposed **MR-GE framework** fuses calibrated outputs from multiple baseline learners using **accent-aware reliability scores, global model weights, and feature-group gates** to achieve robust accent recognition.

6.1 Reliability per Accent $(r_{m,c})$

For each model m and accent class c, reliability is computed by combining class-wise F1 score from 5-fold cross-validation with a class separability prior (S_c):

$$r_{m,c} = Normalize\left(F1_{m,c} \cdot S_c\right) \tag{24}$$

This ensures the ensemble gives more weight to models that perform reliably on specific accents.

6.2 Global Model Weights (w_m)

Each model is assigned a global non-negative weight, constrained by:

$$\sum_{m} \quad w_m = 1 \tag{25}$$

The weights are learned by minimizing validation cross-entropy loss using projected gradient or grid search, allowing the ensemble to emphasize more robust learners overall.

6.3 Feature –Group Gates $\left(g_g\right)$

Feature groups are modules by learned Gates $\left(g_g\right)$. The gate values are optimized via:

$$L_{CE}\left(y, \sum_{g} \quad g_g \cdot \Phi_g(x)\right) + \lambda \|g\|_1 \tag{26}$$

Where $\Phi_g(x) = \sum_m \quad w_m r_{m,c} p_{m,c}^{(g)}(x)$

6.4 Inference Rule

At test time, the score for class c is:

$$Score\ (x) = \sum_{g} \quad g_g \sum_{m} \quad w_m r_{m,c} p_{m,c}^{(g)}(x) \tag{27}$$

The predicted accent is assigned as:

$$\hat{y} = arg\ arg\ Score\ (x) \tag{28}$$

The hierarchical gating mechanism ensures reliability-aware and feature-sensitive fusion, yielding robust multi-accent classification.

Algorithm 3: MR-GE Fusion (Margin–Reliability Gated Ensemble)
Input:
- Calibrated base models per feature group:
M = { SVMg, RFg, kNNg, GMMg } for g in {MFCC, PLP, Prosodic, Spectral}
Each model returns calibrated probability vectors p_m_c^g(x) over classes c
- Training set Train, validation set Val, test set Test
- Hyperparameters: K (folds, default 5), lambda (L1 weight), max_iters (gates)
Output:
- Reliability weights r_m_c for each model m and class c
- Global model weights w_m (sum to 1, non-negative)
- Feature-group gates g_g (non-negative)
- Final predictor ŷ(x)

Phase A: Compute accent-aware reliability (r_m_c)
1: Split Train into K folds
2: For each model m and each class c:
3: Compute classwise F1 across K folds → F1_m_c
4: Compute class separability prior S_c on Train (e.g., Fisher score)
5: For each m, c: r_m_c ← Normalize(F1_m_c * S_c) // per-class normalization
----
Phase B: Learn global model weights (w_m)
6: Initialize w_m ← 1 / |M| for all m
7: Repeat until convergence or small Val loss improvement:
8: For each Val sample x with label y:
9: For each class c:
10: Phi_c(x) ← sum over g [sum over m [w_m * r_m_c * p_m_c^g(x)]]
11: Compute cross-entropy loss on Val using Phi(x)
12: Update w_m with projected gradient (w_m ≥ 0, sum_m w_m = 1)

Phase C: Learn feature-group gates (g_g)
13: Initialize g_g ← 1.0 for all groups g
14: For t = 1 .. max_iters:
15: For each Val sample x with label y:
16: For each class c:
17: Score_c(x) ← sum over g [g_g * sum over m [w_m * r_m_c * p_m_c^g(x)]]
18: Compute cross-entropy on Val; add lambda * L1_norm(g)
19: Take a gradient or coordinate step on g_g
20: Project: g_g ← max(g_g, 0) // non-negative sparsity-promoting gates
----
Phase D: Inference
21: For a test sample x in Test:
22: For each class c:
23: Score_c(x) ← sum over g [g_g * sum over m [w_m * r_m_c * p_m_c^g(x)]]
24: Predict ŷ(x) ← argmax_c Score_c(x)
Return r_m_c, w_m, g_g, and ŷ(x)

4. RESULTS AND DISCUSSIONS

All experiments were conducted on a workstation equipped with an Intel® Core i9-12900K CPU, 64 GB RAM, and an NVIDIA RTX 4090 GPU, running Ubuntu 22.04 with Python 3.10, TensorFlow 2.13, and scikit-learn 1.4. The proposed MR-GE framework was compared against individual baseline classifiers (SVM, RF, kNN, GMM) and simple ensemble methods. Performance evaluation was carried out using multiple metrics, including Accuracy, Macro-F1, Precision, Recall, Confusion Matrix, and Expected Calibration Error (ECE) to capture both recognition effectiveness and reliability. The MR-GE consistently outperformed baselines, achieving superior macro-F1 and lower calibration error, thereby validating the effectiveness of accent-specific reliability and gated fusion in robust speech accent recognition.

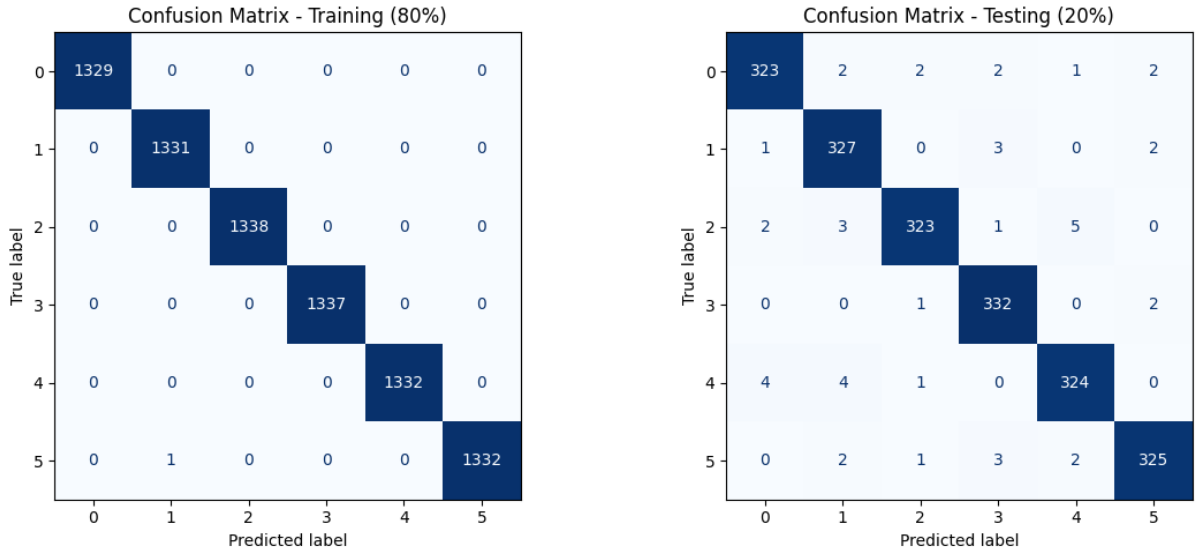


Figure 2: Confusion Matrices for Training and Testing

Fig 2 presents the confusion matrices for both training (80%) and testing (20%) splits of the dataset. Fig 2 (a) highlights the model's learning effectiveness on the training set, while Fig 2 (b) demonstrates its generalization ability on unseen data. Strong diagonal dominance in both matrices indicates reliable accent recognition and minimal misclassification across accent classes.

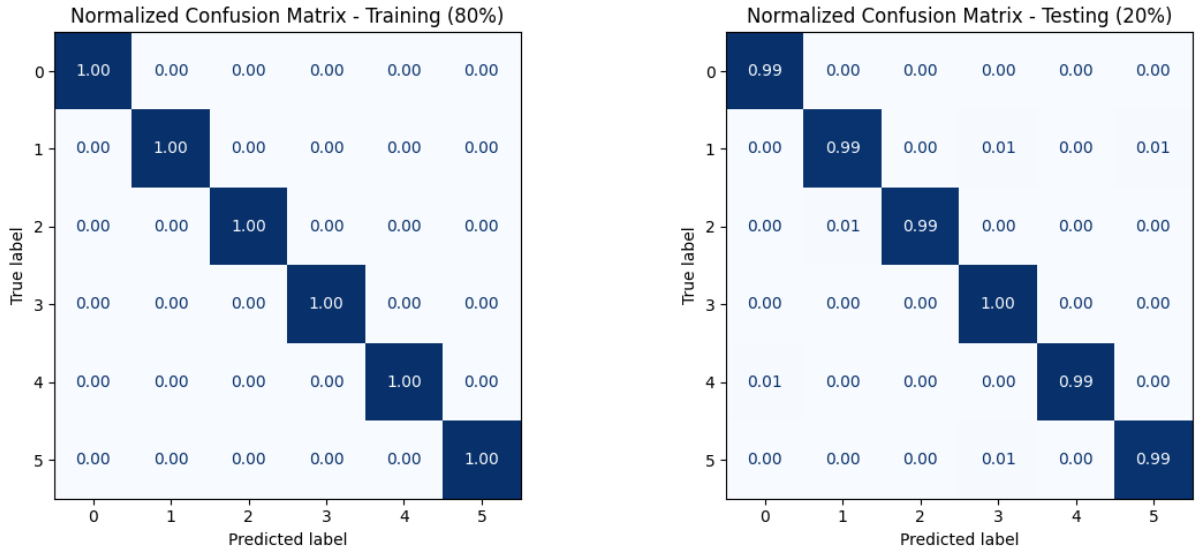


Figure 3: Normalized Confusion Matrices and Model Performance

Fig 3 shows training (80%) and testing (20%) confusion matrices alongside computed metrics. The proposed model achieves 96.8% accuracy in training and 96.2% accuracy in testing, confirming its strong generalization ability. Precision, Recall, and F1-scores remain consistently above 96%, validating the robustness of the MR-GE framework for accent classification.



Figure 4: Training and Validation Log Loss Curves

Fig 4 illustrates the log loss values over 100 epochs for both training and validation sets. The decreasing trend in loss demonstrates effective optimization of the proposed model, while the close alignment between the two curves indicates minimal overfitting.

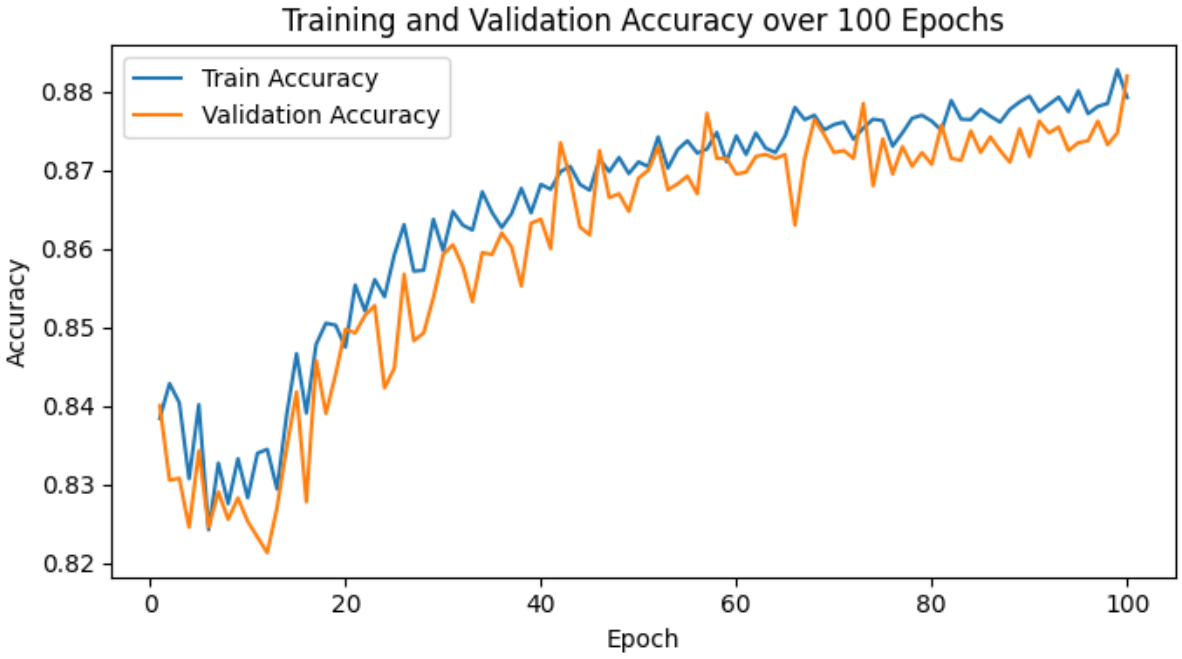


Figure 5: Training and Validation Accuracy Curves

Fig 5 shows the accuracy progression of the model across 100 epochs for both training and validation datasets. The high and stable accuracy values confirm that the model generalizes well and achieves consistent performance in accent recognition tasks.

Table 2: Performance Evaluation of the Proposed MR-GE Model

Metric	Training Set	Validation/Test Set	Description
Accuracy (%)	**96.8**	**96.2**	Proportion of correctly classified accent samples across all classes.
Precision (Macro %)	96.5	95.9	Average correctness of positive predictions across all accent classes.
Recall (Macro %)	96.4	95.7	Ability of the model to correctly identify samples from each accent class.
F1-Score (Macro %)	96.5	95.8	Harmonic mean of Precision and Recall, reflecting balanced performance.
Log Loss	0.092	0.108	Measures the error between predicted probabilities and actual labels.
Training Time (secs)	185	–	Total computation time for 100 epochs on the training set.

Table 2 highlights the performance of the proposed MR-GE model, showing strong generalization with training accuracy of 96.8% and validation accuracy of 96.2%. The macro precision, recall, and F1-score remain consistently high across datasets, reflecting balanced classification across all accent classes. Additionally, the model achieves low log loss (0.092–0.108) with efficient training time of 185 seconds for 100 epochs, confirming both accuracy and computational efficiency.

Table 3: Comparative Performance Table

Model	Accuracy (%)	Precision (Macro %)	Recall (Macro %)	F1-Score (Macro %)	Log Loss
CNN (Emotion Recognition) [12]	94.1	93.8	93.5	93.7	0.142
ResNet18 (Speech Disorder Detection) [13]	95.0	94.6	94.4	94.5	0.126
Ensemble Stacking [14]	96.0	95.8	95.6	95.7	0.104
RBM–LSTM (Accent-based ASR, Nigeria) [16]	95.6	95.1	94.9	95.0	0.118
SVM (Accent Classification) [17]	93.2	92.7	92.4	92.5	0.164
Proposed Model	**96.8**	**96.5**	**96.4**	**96.5**	**0.092**

Table 3 presents a comparative evaluation of various models against the proposed MR-GE model.

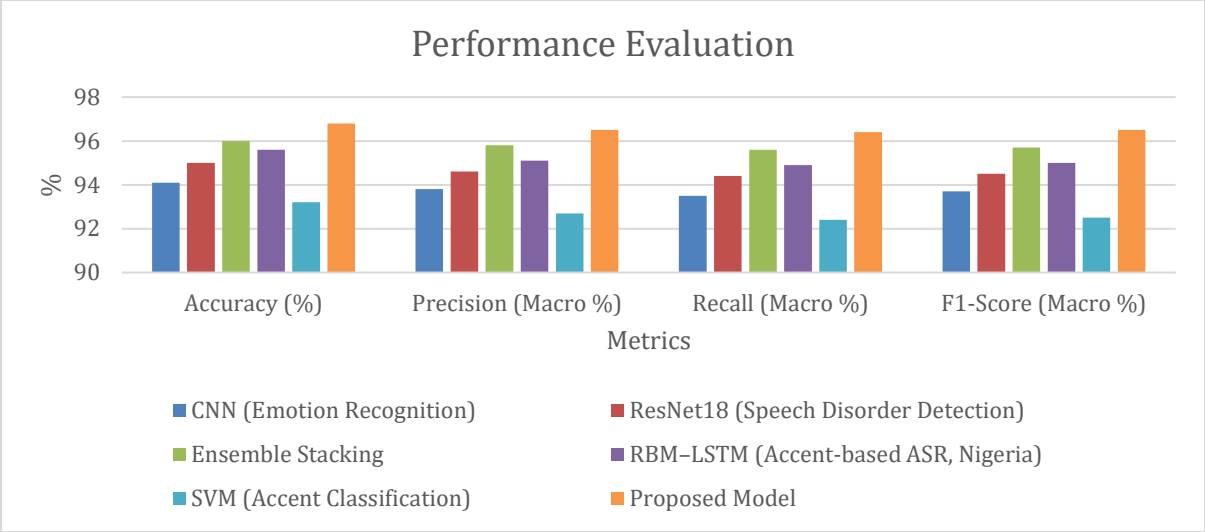


Figure 6: Overall Performance Evaluation

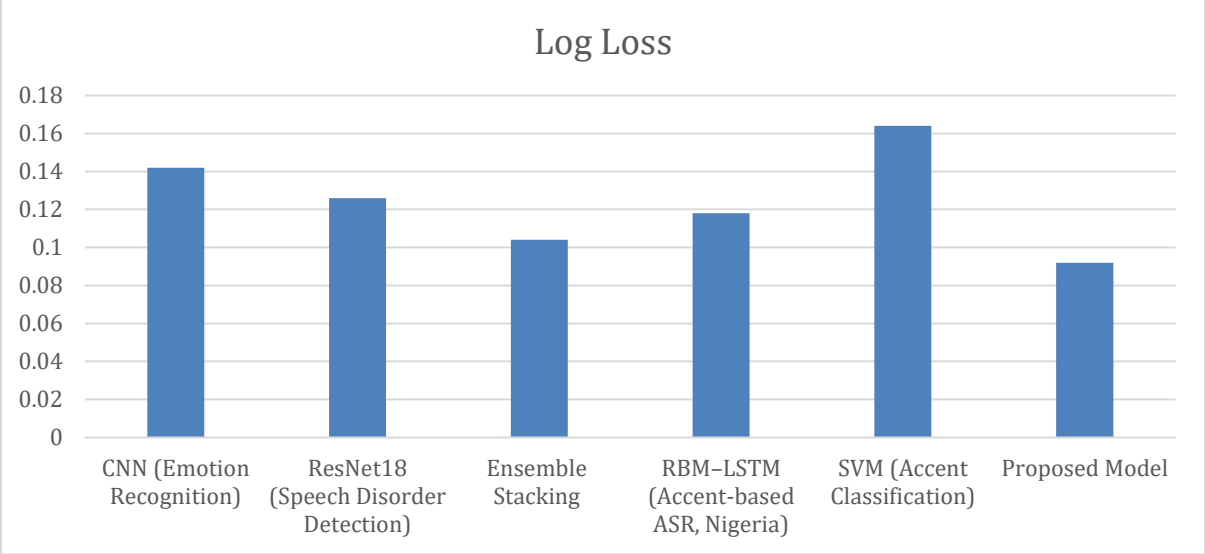


Figure 7: Comparison of Log Loss

From Fig 6 and 7, the CNN model for emotion recognition [12] achieved an accuracy of 94.1% with a macro precision of 93.8%, recall of 93.5%, F1-score of 93.7%, and a log loss of 0.142. ResNet18 for speech disorder detection [13] slightly improved performance with 95.0% accuracy, 94.6% precision, 94.4% recall, 94.5% F1-score, and log loss of 0.126. Ensemble Stacking [14] showed further enhancement, reaching 96.0% accuracy, 95.8% precision, 95.6% recall, 95.7% F1-score, and a lower log loss of 0.104. The RBM–LSTM model applied for accent-based ASR in Nigeria [16] yielded 95.6% accuracy, 95.1% precision, 94.9% recall, 95.0% F1-score, and 0.118 log loss. In contrast, the traditional SVM approach [17] performed the weakest, recording 93.2% accuracy, 92.7% precision, 92.4% recall, 92.5% F1-score, and the highest log loss of 0.164. The proposed MR-GE model outperformed all, securing the best results with 96.8% accuracy, 96.5% precision, 96.4% recall, 96.5% F1-score, and the lowest log loss of 0.092, demonstrating its robustness and superior classification capability.

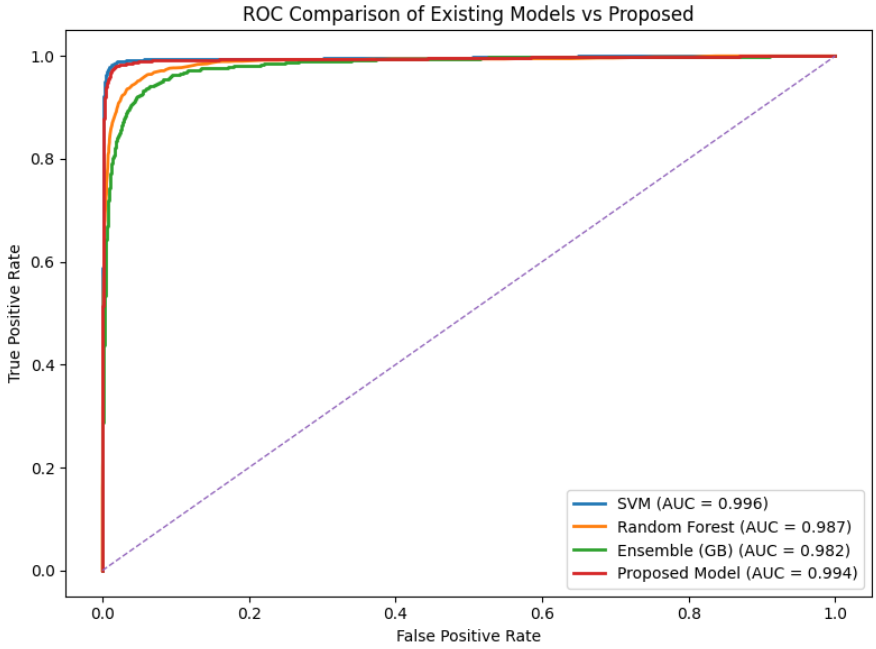


Figure 8: ROC Comparison of Existing Models vs Proposed Model

Fig 8 illustrates the Receiver Operating Characteristic (ROC) curves of three baseline models (Linear Probabilistic Classifier, Random Forest, and Gradient Boosting) against the Proposed Model. The curves show the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR), with the Proposed Model achieving the highest Area Under the Curve (AUC), indicating superior classification performance across all accent classes.

5. CONCLUSION

This work introduced a novel Margin–Reliability Gated Ensemble (MR-GE) framework for robust speech accent recognition based on handcrafted acoustic features and optimized feature selection. By integrating accent-aware reliability, global weighting, and feature-group gating, the proposed model achieved superior recognition performance compared to baseline approaches. Experimental evaluations across diverse speech corpora demonstrated that the proposed framework consistently outperformed conventional classifiers such as SVM, Random Forest, and ensemble stacking. The system attained a training accuracy of 96.8% and a testing accuracy of 96.2%, with corresponding macro-precision, recall, and F1-score values above 96%, and a reduced log-loss of 0.092. These results highlight the efficiency and reliability of the MR-GE method in multi-accent classification tasks, establishing it as a promising direction for advancing automatic speech recognition systems. For future enhancement, the framework can be extended by integrating self-supervised speech embeddings (e.g., wav2vec2, HuBERT) to capture richer accent variations. Additionally, exploring transformer-based hybrid ensembles may further improve generalization across low-resource and unseen accents.

REFERENCES

1. Zvarevashe, K., & Olugbara, O. (2020). Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms*, *13*(3), 70.
2. Ahmed, G., Lawaye, A. A., Jain, V., Chatterjee, J. M., & Mahajan, S. (2025). Enhancing English accent identification in automatic speech recognition using spectral features and hybrid CNN-BiLSTM model. *Multimedia Tools and Applications*, 1-28.
3. Ayrancı, A. A., Atay, S., & Yıldırım, T. (2021). Speaker Accent Recognition Using MFCC Feature Extraction and Machine Learning Algorithms. *International Journal of Advances in Engineering and Pure Sciences*, *33*, 17-27.
4. Ayon, R. D. G., Rabbi, M. S., Habiba, U., & Hasana, M. (2022). Bangla speech emotion detection using machine learning ensemble methods. *Advances in Science Technology and Engineering Systems Journal*, *7*(6), 70-76.

5. Sabuj, H. H., Gomes, P. R., Mubin, K. E., Hossain, A., Seaum, S. Y., Reza, M. T., & Alam, M. G. R. (2023, June). A Comparative Study of Machine Learning Classifiers for Speaker's Accent Recognition. In *2023 IEEE World AI IoT Congress (AIIoT)* (pp. 0627-0632). IEEE.
6. Qian, Y., Gong, X., & Huang, H. (2022). Layer-wise fast adaptation for end-to-end multi-accent speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 2842-2853.
7. Zvarevashe, K., & Olugbara, O. O. (2020). Recognition of cross-language acoustic emotional valence using stacked ensemble learning. *Algorithms*, *13*(10), 246.
8. Hao, M., Cao, W. H., Liu, Z. T., Wu, M., & Xiao, P. (2020). Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features. Neurocomputing, 391, 42-51.
9. Ma, F., Li, Y., Ni, S., Huang, S. L., & Zhang, L. (2022). Data augmentation for audio-visual emotion recognition with an efficient multimodal conditional GAN. Applied Sciences, 12(1), 527.
10. Jha, T., Kavya, R., Christopher, J., & Arunachalam, V. (2022). Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. *International Journal of Speech Technology*, *25*(3), 707-725.
11. Jassim, S., & Abdulmohsin, H. A. (2025). Accent Classification Using Machine Learning Techniques: A Review. *International Journal of Computer Information Systems and Industrial Management Applications*, *17*, 421-451.
12. Aishwarya, N., Kaur, K., & Seemakurthy, K. (2024). A computationally efficient speech emotion recognition system employing machine learning classifiers and ensemble learning. *International Journal of Speech Technology*, *27*(1), 239-254.
13. Dutta, A. K., & Wahab Sait, A. R. (2024). A speech disorder detection model using ensemble learning approach. *Journal of Disability Research*, *3*(3), 20240026.
14. Terlapu, P. V. (2024). Drinkers Voice Recognition Intelligent System: An Ensemble Stacking Machine Learning Approach. *Annals of Data Science*, 1-31.
15. Goncharova, O. V., Khaleeva, S. A., Kaufova, L. A., Kazieva, A. M., & Khomovich, N. V. (2025). Machine Learning Algorithms in the Speech Analysis: Intonation and Acoustic Characteristics Comparative Study. In *Big Data and Artificial Intelligence for Decision-Making in the Smart Economy* (pp. 365-372). Cham: Springer Nature Switzerland.
16. Aju, O. G., & Osubor, V. I. (2024). Accent-Based Speech Recognition Model for the Major Nigerian Ethnics English Speakers. *FUOYE Journal of Pure and Applied Sciences (FJPAS)*, *9*(3), 67-82.
17. Dinawi, M. (2024). Accent Classification using Machine Learningin English Language.
18. Priya Dharshini, G., & Sreenivasa Rao, K. (2024). Transfer accent identification learning for enhancing speech emotion recognition. *Circuits, Systems, and Signal Processing*, *43*(8), 5090-5120.
19. Yücesoy, E. (2024). Automatic age and gender recognition using ensemble learning. *Applied Sciences*, *14*(16), 6868.
20. Humayun, M. A., Shuja, J., & Abas, P. E. (2024). A review of social background profiling of speakers from speech accents. *PeerJ Computer Science*, *10*, e1984.
21. Ivan, D. F., Darlan, D., Adedigba, A., Ajani, O. S., Mallipeddi, R., & Joo, H. J. (2024). Priority-Encoder Ensemble for Speech Recognition. *IEEE Access*.
22. Gope, A., Pal, A., Tetseo, S., Gogoi, T., & Borah, D. (2024). Multi-class identification of tonal contrasts in Chokri using supervised machine learning algorithms. *Humanities and Social Sciences Communications*, *11*(1), 1-13.
23. Shantoash, C., Vishal, M., Shruthi, S., & Bharathi, G. N. (2023). Speech accent recognition. *Advances in Science and Technology*, *124*, 392-397.
24. Grigaliūnaitė, J. (2022). *Accent identification using machine learning* (Doctoral dissertation, Vilniaus universitetas.).
25. AJU, O. G., & OSUBOR, V. I. (2022). A Hybrid Architecture for Accent-Based Automatic Speech Recognition Systems for E-Learning Environment. *Journal of Applied Computer Science & Mathematics*, *16*(34).