SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I5P124

Volume 13 Issue 5, 301-311, May 2026

Original Article

Deep Convolutional Framework for Speaker Identification from Emotional Speech Using Multi-Domain Acoustic Feature Fusion

Rupali Khaklary[1], Nabankur Pathak[2]

[1,2]*Krishna Kanta Handiqui State Open University, Guwahati, Assam, India.*

[1]*Corresponding Author : khaklaryrupali9@gmail.com.*

Received: 20 February 2026 Revised: 20 March 2026 Accepted: 22 April 2026 Published: 27 May 2026

Abstract *- Speaker identification from emotional speech is challenging due to variations in acoustic characteristics introduced by different emotional states, which often affect speaker-specific information. This paper introduces a profound convolutional network for speaker identification through multi-domain acoustic features from a custom Bodo dataset. Complementary time–frequency representations, including Mel-Frequency Cepstral Coefficients (MFCC), Log-Mel spectrograms, and chroma features, are combined to capture spectral, perceptual, and harmonic characteristics of speech signals and hence enhance speaker discrimination under emotional variability. The 2-dimensional Convolutional Neural Network is used to acquire hierarchical feature representations to classify multi-class speakers. A Baseline system with MFCC features is created and then measures the effectiveness of feature fusion based on alternative combinations of features. Data augmentation methods, such as the addition of white noise and time stretching, are applied to the training dataset to enhance the generalization and robustness of models. Experimental findings show that the proposed multi-feature fusion approach outperforms with the fusion of MFCC and Log-Mel spectrogram representations, achieving an identification accuracy of 94.53%. The results suggest that the speaker separability is greatly enhanced, and a convolutional architecture is also computationally efficient by incorporating complementary acoustic features.*

Keywords *- Convolutional Neural Network, Deep Learning, Feature Fusion, Spectral features, Speaker Identification.*

1. Introduction

A Speaker Identification (SI) system plays a significant role in biometric security that enhances user security and authentication based on voice characteristics. SI is a highly flexible biometric modality that is essential for ensuring access control, surveillance, and identity checking, where authentication is required, and is an ideal and easily customizable solution for many platforms, including telephones, computers, and mobile phones [1, 2]. Instead of using conventional codes, like passwords or PINs, voice-based authentication relies on the individuality of the vocal characteristics and provides a more secure and user-friendly alternative [3].

Conventional SI systems were based on handcrafted acoustic features and statistical modeling techniques, like GMM-UBM and i-vector representations, usually trained together with classifiers like SVM or PLDA [4-6]. However, these are highly manual in feature design and lack the power to capture complex features of the speaker. These limitations motivated the use of deep learning-based features rather than traditional models, where Convolutional Neural Networks (CNNs) architectures learn speakers' unique voice feature characteristics directly from the given input features [7-10]. Deep embedding-based models such as x-vector, ECAPA-TDNN, and ResNet learn highly discriminative speaker embedding with large-scale data and excel in performance over traditional models [11-13]. Recently, self-supervised learning frameworks such as wav2vec 2.0, HuBERT, and WavLM have been developed using large volumes of unlabeled speech data, especially in the context of low-resource or domain-mismatched settings [14-16]. However, these models are usually pretrained on high-resource languages, and this restricts their generalization capability to low-resource languages like Bodo.

Studies [17, 18] reported that low-cost deep models are effective in low-resource tasks due to reduced computation and overfitting with hybrid approaches. Spectral features are rich in phonetic and speaker-specific information that can be utilized in biometric identification tasks, and a Deep Learning (DL) model can use these to identify speakers with high precision [19-21]. Experimental studies [22-27] have proved that the use of MFCC-based features improves identification

performance and lowers error rates, gains robustness and recognition rates in noisy environments. CNNs are effective at capturing local spectral patterns and hierarchical representations of the time-frequency spectrums [26, 28-31], which further strengthen the model's performance. Moreover, studies in [32-34] have proven that complementary acoustic feature representations are more effective and reliable than single-feature representations in discriminating speakers' voice characteristics. Researchers have presented various research frameworks for SI systems throughout different languages using both traditional models and DL-based approaches [35-39]. However, several existing studies have concentrated on resourceful languages, but the low-resource Bodo language is still unexplored. Although prior studies [25, 38] have explored CNN-based speaker approaches for SI and ER in other low-resource languages, the models were isolated and not within the context of Bodo. Moreover, the influence of emotional variability on SI has not been addressed in these studies.

This defines an important limitation that emotional variations significantly affect acoustic characteristics and lower the system performance. Comparatively, this paper examines SI in Bodo speech containing emotionally expressive information and a methodical examination of spectral feature fusion. To make it clear at a glance, a structured comparison of existing studies and to highlight the research gap, a summary is presented in Table 1. To address this gap, the proposed research will establish an SI model that is Bodo emotional speech-specific. This work aims to develop an effective and robust model that can identify speakers from their variant emotional speech and, hence, mitigate the constraints of the low-resource setting of Bodo and thereby contribute to diversity in speech technology. The research sets up an MFCC-based CNN model as a baseline and compares spectral feature fusion under the same configuration setting by evaluating MFCCs and their combinations with the mel-spectrogram and chroma. Features to differentiate between speakers in emotional Bodo speech. Figure 1 is the block diagram of the proposed model, presenting the pipelines that will be implemented in the model.

The main contributions of this study are summarized as follows:

- Formulation of a speaker identification framework of Bodo speech using complementary spectral features, MFCC, mel-spectrogram, and chroma, and their combinations with a single CNN-based system.
- The systematic study of spectral feature fusion is performed to determine the contribution of the various representations towards speaker discrimination in expressive speech having emotional overtones.
- Introducing a well-organized design with simple pre-processing, regulated augmentation, and evaluation without leakage of data.
- The experimental findings prove that fusion of MFCC and mel-spectrogram features enhances the identification performance rate as compared to single-feature representations and conventional baselines under the unified setting.
- The study offers an empirical insight into speaker recognition of the low-resource language Bodo, in which the large-scale embedding-based training is unattainable.

Including the introduction section, the paper is structured with 4 sections where Section 2 describes the dataset and approaches used in the study, Section 3 summarizes the experimental results and discusses the analysis, Section 4 offers a brief conclusion, and outlines the future work.

Table 1. Comparative analysis of existing studies that are relevant to the speaker identification framework and the current study across key research dimensions

Study	Model	Emotion considered	Low-resource Language	Feature fusion	Limitation
[20]	DNN (Fusion)	X	X	✓ (RW and GTCC feature)	High complexity
[25]	CNN	X	✓ (Indonesian)	✓ (DWT + MFCC)	No emotional speech
[27]	CNN	X	X	X (MFCC)	No emotional or low-resource focus
[33]	Hybrid	X	X	✓	No emotional modeling
[36]	CNN	✓	X	X (MFCC)	No language-specific modeling
[37]	ML-based	✓	X	✓ (Spectral)	Not focused on SI
[38]	HMM	X	✓ (Urdu)	X (MFCC)	No deep learning, no emotion
[39]	Classical	X	✓ (Turkish)	X (Acoustic)	Limited scalability
[40]	ML-based	X	✓ (Arabic)	X (Phoneme-based)	No emotional modeling
This Study	Deep CNN	✓	✓ (Bodo)	✓ (MFCC, Chroma and Mel-spectrogram)	Limited emotional speech data with no cross-language or cross-corpus validation.

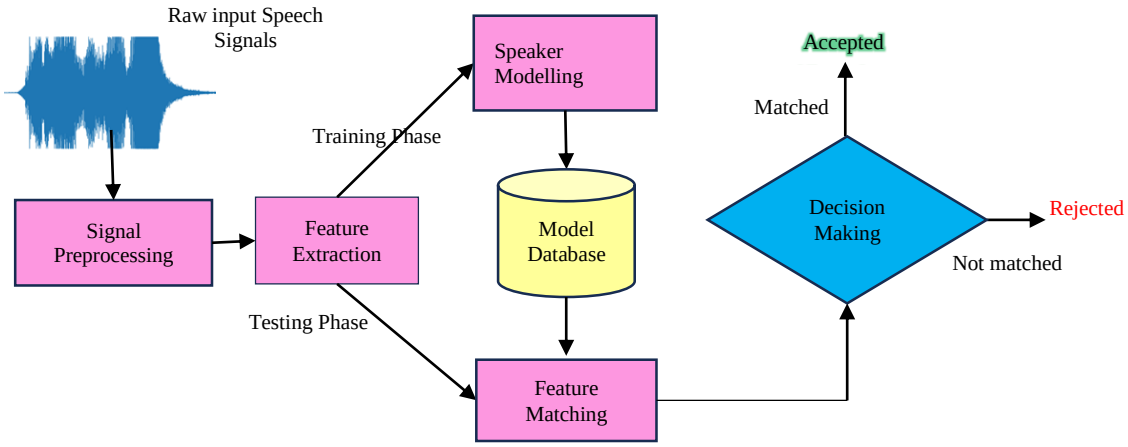

Fig. 1 Block diagram of the proposed speaker identification pipelines

2. Materials and Methods

2.1. Dataset Description

The data used in the study is a locally collected Bodo speech corpus, which was obtained by means of the native speakers of Bodo with the open-source software Audacity and a Condenser microphone SF-666 in a mono channel and 16kHz frequency. The corpus is a balanced dataset including 3960 audio WAV files in six emotional (neutral, happy, sad, angry, fear, and surprise) states of 22 speakers (11 men and 11 women) who equally contributed 180 utterances on a standard script comprising 15 sentences. The audio was recorded in an ordinary room setting where the average duration of the recorded utterances is 2.15 seconds.

2.1.1. Dataset Preparation

For evaluation purposes, the complete dataset of 3960 samples is divided into two subsets: training plus validation and testing in the ratio 75:25 at the utterance level. Again, 20% of the first subset is further separated to form the validation set for model tuning. Keeping the speaker set constant across all subsets, the utterances are assigned to all sets in a mutually exclusive manner to prevent data leakage. The test is based on a closed-set speaker identification protocol, where the model is trained and tested on the identical speaker set to determine its capability and identify the speaker in unseen utterances at different emotional settings.

2.2. Preprocessing

Noise and silence removal were performed manually using the open-source software Audacity. Then, audio signals are resampled to a uniform sampling rate of 16KHz, amplitude is normalized and scaled to a common duration by zero-padding shorter signals and truncating longer signals to have a common input length for feature extraction. The remaining techniques are implemented as follows-

2.2.1. Feature Scaling (Standardization)

A standardization method is used to put the data of different scales into a similar range. It alters the data to an average of 0 and a standard deviation of 1. The standard-scaled value X_{scaled} of a feature X with a mean μ is obtained by the formula (1).

$$X_{scaled} = \frac{X-\mu}{\sigma} \quad (1)$$

This role homogenizes the quality by eliminating the average and normalizing it to unit variance. This guarantees consistency in the process of data preprocessing, thereby helping the model to be trained, evaluated, validated, and tested in the right way without misleading the integrity of the process.

2.2.2. Data Augmentation

Due to the scarcity of available samples, the training subset was augmented to bring variation in data and hence improve model generalization and robustness against variations in input conditions. The validation and testing sets contain original recordings so that evaluation is performed on unseen and unaltered speech data. The following two methods are applied for this purpose-

Time Stretching

This refers to a method of varying the speed of an audio signal without altering the pitch. Given a speech signal $x(t)$, the time-stretched signal $x'(\alpha t)$ is obtained by (2),

$$x'(t) = x(\alpha t) \quad (2)$$

where α is the stretching factor. When $\alpha > 1$, the signal is accelerated, and it is decelerated when $\alpha < 1$.

Adding White Noise

White noise was added to the original signal to increase the variability in the input data, which is limited. To a signal S, a white noise, N, is injected using the formula (3),

$$S_{noisy} = S + \alpha \cdot N \quad (3)$$

where α is a scaling factor that controls the amount of noise added to the signal.

The original training set, consisting of 3260 samples, is thus augmented to a threefold (3×) expansion yielding a total of 9780 training samples.

2.3. Feature Extraction

The preprocessed signals are divided into overlapping frames with the help of a Hamming window of length 25 ms and a hop length of 10ms. Then each frame is subjected to Short-Time Fourier Transform (STFT) using a 512-point FFT in order to minimize spectral leakage. The resultant time frequency representation is now used to compute MFCC, log-mel spectrogram, and chroma features. Let the windowed speech frame of length N be x[n], its frequency-domain representation is obtained through the Discrete Fourier Transform (DFT) as in (4)

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \tag{4}$$

and the corresponding power spectrum is computed as (5)

$$P[k] = |\, X[k]\|^{\,2} \tag{5}$$

where k is the frequency bin index. The power spectrum is used to extract subsequent features.

2.3.1. Log-Mel Spectrogram Extraction

The power spectrum is then filtered by a triangular filterbank that is separated by the mel scale. The relationship of the mel scale frequency (f_{mel}) and linear frequency f is defined by (6).

$$f_{mel} = 2595\, log_{10}\,(1 + \tfrac{f}{700}) \tag{6}$$

The output (energy) (X_k) of the k- th filter is obtained by (7).

$$S_k(t) = \textstyle\sum_m P_t[m] H_k[m] \tag{7}$$

Where $P_t[m]$ denotes the power spectrum and $H_k[m]$ is the magnitude response of the *k*-th mel filter, and m is the index of the frequency bin within the filter range. The output, Mel-filter bank energies, is then logarithmically transformed to resemble the audibility of human ear perception. It is performed by (8).

$$Log - Mel_k(t) = \log(S_k(t)) \tag{8}$$

This produces a two-dimensional log-Mel spectrogram, which describes the distribution of spectral energy across time and perceptual frequency bands. The temporal mean and standard deviation of each Mel-frequency band were calculated to get fixed-length representations that can be used in machine learning models as (9).

$$\mu_k = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} Log - Mel_k(t)\ ;$$

$$\sigma_k = \sqrt{\tfrac{1}{T}\textstyle\sum_{t=1}^{T}(Log - Mel_k(t) - \mu_k)^2} \tag{9}$$

where T is the number of frames. The concatenation of these statistics forms the final log-Mel feature vector.

2.3.2. MFCC Feature Extraction

MFCC features provide a compact representation of the spectral envelope of speech signals. The DCT is used to decorrelate the log-Mel coefficients (log-Mel filter bank energies) and obtain the MFCCs by (10).

$$MFCC_i = \textstyle\sum_{k=1}^{K} Log - (Mel_k) Cos[\, i.\left(k - \tfrac{1}{2}\right).\tfrac{\pi}{K}\,] \tag{10}$$

where X_k is the magnitude spectrum of the speech signals, K is the number of mel filter banks, and i is the index of MFCC coefficients [40].

In this way, MFCC features were extracted using 128 Mel filter banks with a frame length of 25 ms and a frame shift of 10 ms. In order to maintain more spectral detail in the feature representation, 40 MFCC coefficients were retained to be further analyzed, including higher-order cepstral coefficients, excluding delta and delta-delta coefficients.

2.3.3. Chroma Feature Extraction

Chroma features describe the spectral energy allocation of the 12 distinct pitch classes (semitones) of a musical octave, which are independent of absolute pitch or octave information. These are used to capture the harmonic content of speech. The spectral energy is projected onto chroma bins by grouping frequency components belonging to the same pitch class. The chroma vector, $chroma_i$ for the i-th pitch class is calculated by (11).

$$Chroma_i(t) = \textstyle\sum_{k \in C_i} |X_t[k]| \tag{11}$$

Where C_i is the set of frequency bins corresponding to the i-th pitch class, and $|X_t[k]|$ is the magnitude of the spectrum at frequency bin k for frame *t* [41].

The spectral energy across all octaves is then summarized in a single pitch class representation to produce 12-dimensional chroma vector per frame. To derive a fixed length feature representation, the mean and standard deviation of each chroma constituent were computed over all the frames, and these were taken as input features to further modeling.

Table 2. Summary of extracted features to be employed for model training

Feature	Representation	Dimension
Log-Mel Spectrogram	Mean + Standard deviation over time	2 x 128 =256
MFCC	Mean + Standard deviation over time	2 x 40 = 80
Chroma	Mean + Standard deviation over time	2 x 12 = 24

Thus, the resultant concatenated feature vector has a dimension of 360 as summarized in Table 2, which proceeds to the training model.

2.4. Model Architecture and Setup

The proposed SI model was constructed using 2-dimensional CNNs for multi-class classification, as shown in Figure 2**.** The architecture is configured to acquire hierarchical feature representations out of the time-frequency representation of input features derived from the extracted features representation. All convolutional and intermediate fully connected layers based on Real-valued numbers use a Rectified Linear Unit (ReLU) activation function to add nonlinearity and increase learning capacity.

The CNN with MFCC is set as the baseline system. Further, the same architecture is tested on Log-Mel spectrogram, chroma features, as well as combinations of both to examine their impact on feature representation. The proposed model would be associated with the CNN being trained on the combined feature representation, which enables the network to utilize complementary spectral and harmonic data given the same architectural conditions.

The network is made up of four convolutional layers through which the input feature representation is processed. The initial convolutional layer applies 256 filters and obtains the low-level spatial features, and then the max-pooling operation is used to reduce the dimensionality. Then, the next convolutional layer comprising 256 filters further refines the extracted feature representations, followed by another max pooling. The third convolutional layer uses 128 filters to make more abstract-level learnings, after which there is max-pooling and a dropout layer to minimize overfitting. A fourth convolutional layer consisting of 64 filters gets compact discriminative features, and the final max-pooling stage comes after it.Then the resulting feature maps are flattened into a single 1-dimensional feature and passed to a fully connected layer of 32 neurons to accomplish nonlinear feature integration. Followed by the last stage of classification, a dropout layer is implemented. The analysis layer has 22 neurons, which are configured with a softmax activation function to give out class probability distributions. For training, the proposed model was set up with 100 epochs using the Adam optimizer with categorical cross-entropy loss and a batch size of 32. A dropout rate of 0.3 was also applied to reduce overfitting. The initial learning rate was 0.001 and adjusted by a ReduceLROnPlateau strategy on the basis of training loss.

2.5. Evaluation Matrices

Several measures were used to test the proposed model, including accuracy, confusion matrix, precision, recall, and F1-score.

3. Experimental Findings and Analysis

3.1. Performance Comparison

Table 2 shows the classification performance attained with the unified CNN model when varying feature representations were used.

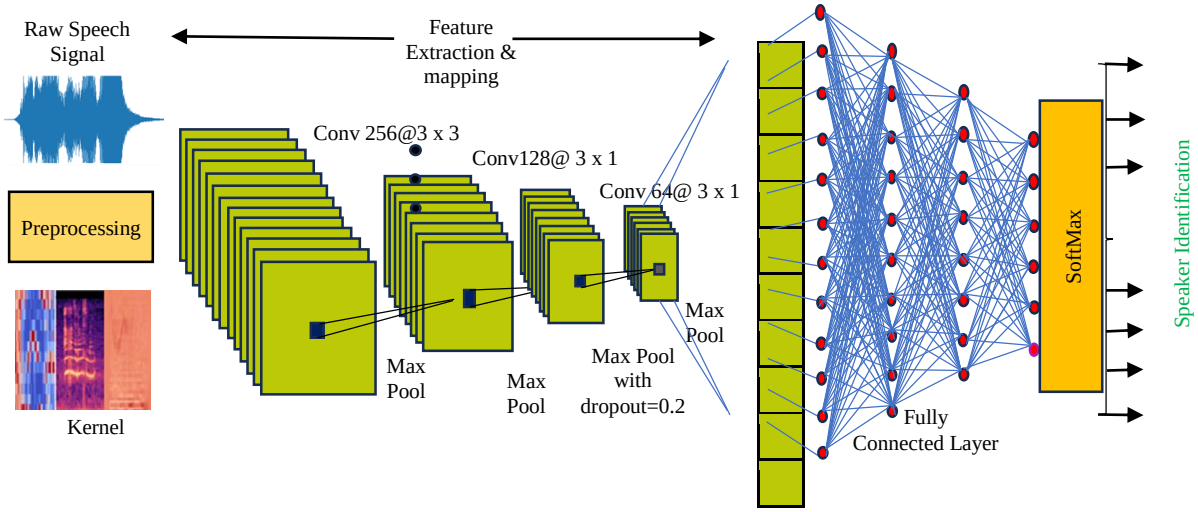

Fig. 2 Model Architecture of the proposed CNN-based Speaker Identification Model

Table 2. Model's summary results across different feature combinations evaluation metrics

Feature Combinations	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
MFCC	87.37	87.40	87.50	87.59
Chroma	69.76	69.77	69.86	69.59
Mel-Spectrogram	85.25	85.54	85.40	85.31
MFCC + Chroma	86.23	86.40	86.36	86.31
MFCC + Mel-Spectrogram	94.53	94.62	94.53	94.53
Chroma + Mel-Spectrogram	84.28	84.45	84.50	84.45
MFCC + Chroma + Mel-Spectrogram	91.21	91.40	91.18	91.22

The baseline MFCC-based system exhibited a classification accuracy of 87.37% and has a reliable performance in speaker identification. The chroma feature representation alone gave a relatively lesser performance, meaning that harmonic information alone is inadequate in giving strong speaker discrimination.

Mel-spectrogram representation outperformed MFCC and proved the significance of time-frequency energy details. Further, the combination of feature representations increased the performance since complementary data in the various feature domains allowed the CNN to learn features better. The MFCC and Mel-spectrogram combination proved to perform the best, with the highest accuracy of 94.53 % and the highest values of precision, recall, and F1-score. This enhancement indicates that spectral envelope representations of MFCC and detailed time-frequency representations of Mel-spectrogram characteristics furnish very complementary representations of speakers.

The three-feature fusion (MFCC + Chroma + Mel-spectrogram) also performed well, but it is shown that chroma feature addition with other features declines the performance rather than improves it in each case. Perhaps this is because of redundancy or less significant discriminant contribution in the case of the given task. MFCCs and log-mel are those features that highlight the spectral envelope of vocal tract properties and thus are very discriminative to speaker identity [42, 43]. Conversely, chroma features are an octave-folded pitch class data originally intended to analyze music [41], thus overriding speaker-specific harmonic spacing. Moreover, tonal and emotional speech causes a notable difference in pitch variation, which enhances intra-speaker variation and decreases the accuracy of pitch-class depiction of a speaker [44, 45].

In general, the findings suggest that the proposed method is effective as it enables the use of the same network architecture with proper feature fusion that enhances classification performance.

The confusion matrix of the model that achieved the highest performance is shown in Figure 3, and Figure 4 presents the training vs validation graph of the lowest performance with chroma feature and the highest performance.with feature fusion of all three features. Figure 5 presents speaker-wise recognition reported by the model. The comparable Accuracy, Precision, Recall, and F1-score values across different combinations of speech features utilized for evaluation are presented in Figure 6. The consistency of the performance outcomes, as depicted in the graph, indicates that the model is good at handling false positives and negatives and makes dependable predictions of all classes.

3.2. Discussion

Because of the lack of standardized benchmarks for the Bodo language and differences in experimental conditions among the literature, technically, it is impractical to compare directly on a quantitative basis with previously reported methods. Nevertheless, the results obtained in this work provide clear evidence of the effectiveness of the proposed framework. The experimental findings prove that the model successfully learns hierarchical patterns from time-frequency representation features and is able to effectively discriminate the voices of the speakers.

CNNs are particularly effective at capturing structural relationships within spectrogram-based inputs, which contributes to improving the classification performance [29, 46, 47]. The result also indicates that the combination of multiple feature types improves resistance to variation within the same speaker since they encode multiple acoustic characteristics at the same time.

The combined representation of this model enables it to differentiate more delicate speaker-specific attributes that could not be represented by an individual feature. Earlier existing studies have also suggested that deep CNN-based models can enhance generalization ability with multi-feature representations and facilitate more effective learning of abstractions [34, 48, 49].

Overall, the performance gains of the model can be characterized by the combined effect of complementary feature representations, data augmentation strategies, and an efficient deep learning architecture framework.

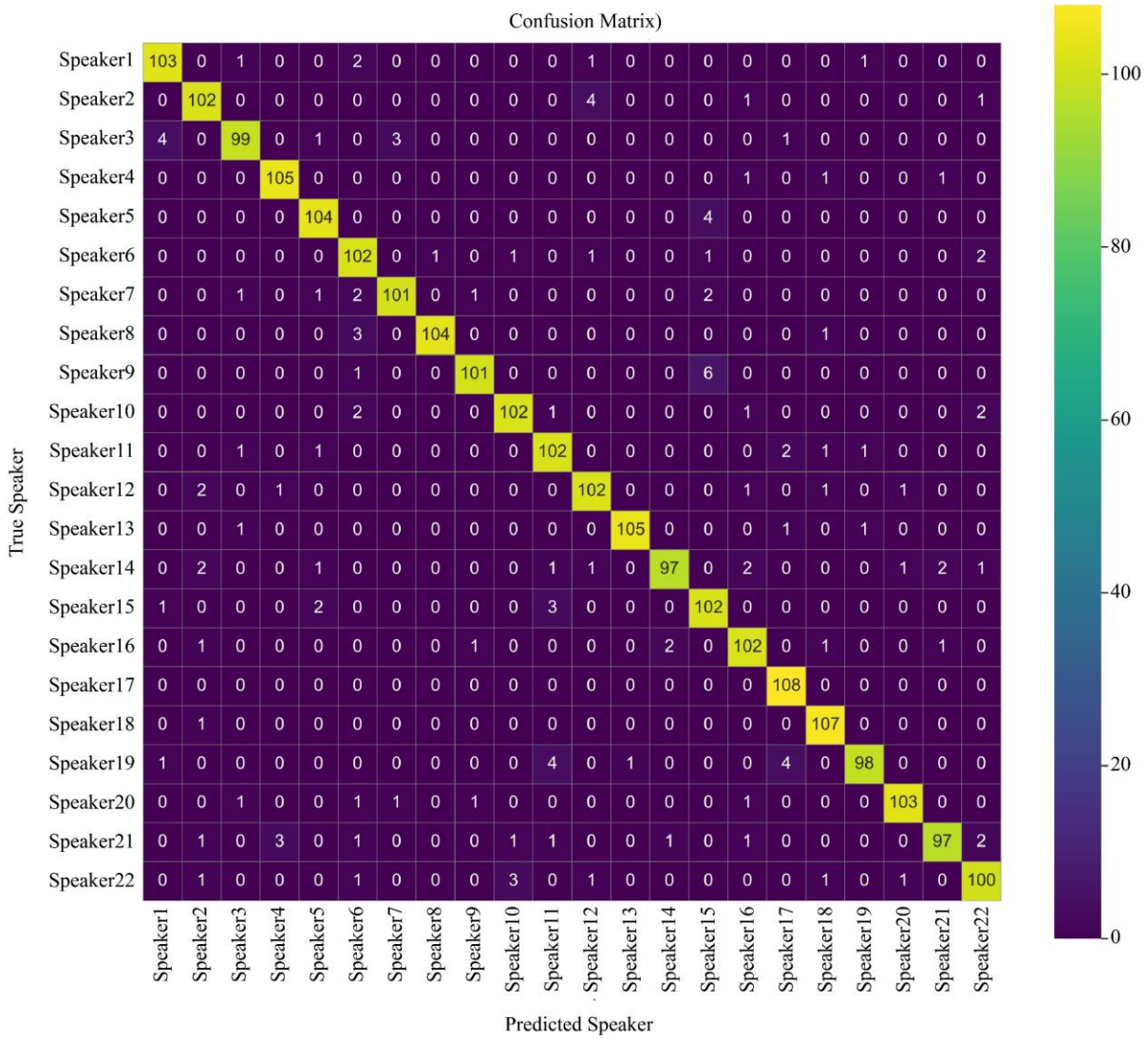

Fig. 3 Confusion matrix of the best performance model visualizing prediction results by comparing the predicted class label, the x-axis (Predicted Speakers) with the y-axis, the true label (Actual Speakers)

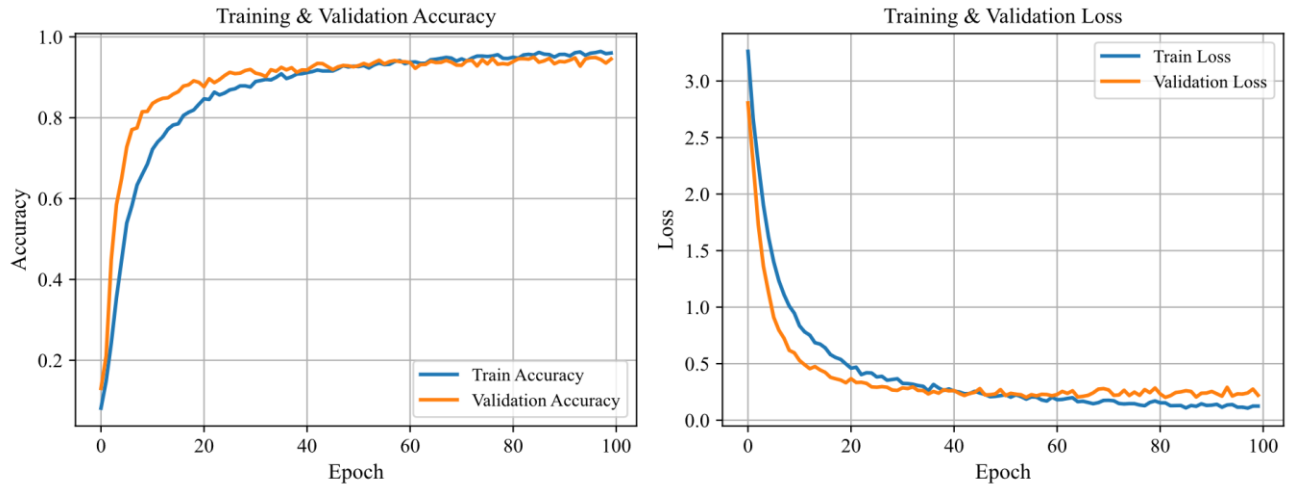

Fig. 4 Display of model performance presenting training vs validation accuracy (left) and training vs validation loss (right)

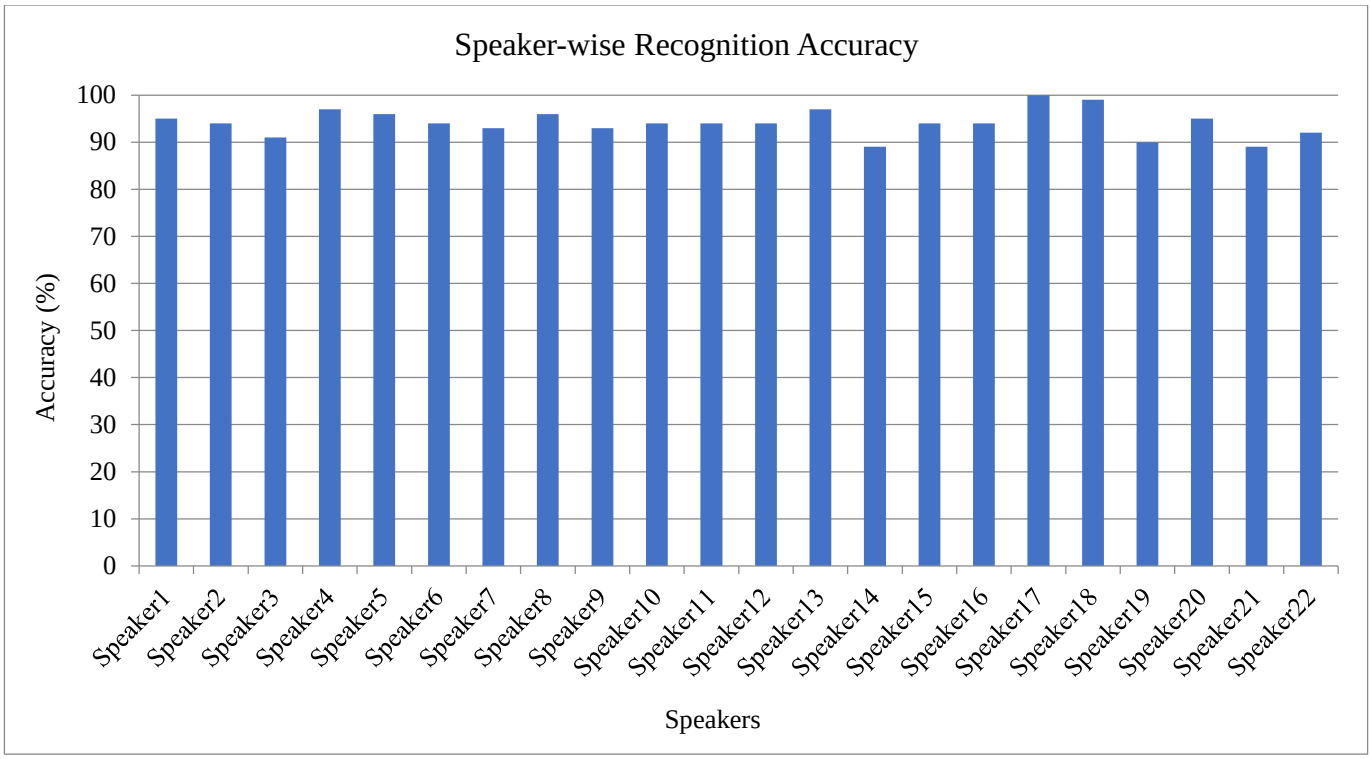

Fig. 5 Graph of speaker-wise classification accuracy (%) indicating consistent results across all speaker classes

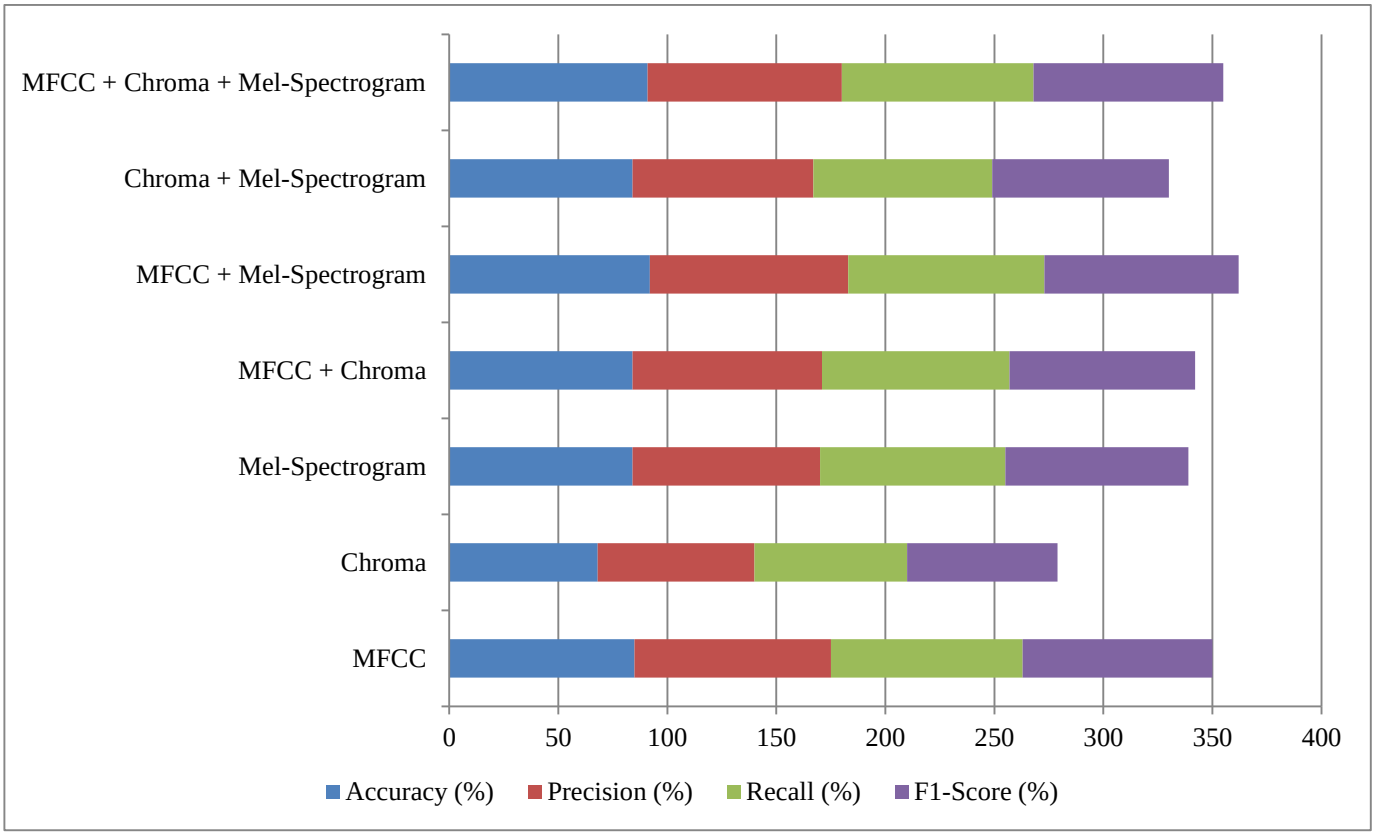

Fig. 6 Graphical visualization of Accuracy, Precision, Recall and F1-score matrices of the model in different combinations of MFCC, Chroma and Mel-spectrogram

4. Conclusion

This study presents a CNN-based SI framework for the Bodo language using multiple time-frequency. representations of speech signals, including how the influence of MFCC, Mel-spectrogram, and chroma affects the classification performance on the basis of individual and combination within the same unified architecture. The experimental findings show that the combination of MFCC and Mel-spectrogram yields the best performance among the evaluation settings, as demonstrated through the ablation study. The results show that the proper choice of feature representation combinations is more efficient than simply increasing the dimensionality of features, the proposed approach is also effective in capturing speaker-specific properties and is consistent in the performance across speakers, i.e., it is strong in multi-class classification tasks. Moreover, data augmentation techniques are also integrated, which in turn helps the generalization to be improved, which is particularly critical in low-resource settings (when it comes to the Bodo language). The findings indicate that the suggested framework can be used as an efficient basis to build strong speech-based systems in languages that are underrepresented. The future work is on attention-based architectures and adaptive feature fusion plans in larger datasets so as to further enhance model performance and scalability.

Conflict of interest

The authors state that they have no financial conflicts or personal relationships that could have affected the results or interpretation of this study.

Funding Statement

This research received no specific grant or support from any funding agency in the public, commercial, or non-profit organizations.

References

[1] Georg Heigold et al., "End-to-end Text-dependent Speaker Verification," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, pp. 5115-5119, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[2] Muhammad Abdul Basit, Chanjuan Liu, and Enyu Zhao, "SDI: A Tool for Speech Differentiation in user Identification," *Expert Systems with Applications*, vol. 243, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Tomi Kinnunen, and Haizhou Li, "An Overview of Text-independent Speaker Recognition: From Features to Supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12-40, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[4] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[5] Patrick Kenny, "Bayesian Speaker Verification with Heavy-tailed Priors," *Odyssey Speaker and Language Recognition Workshop*, 2010. [Google Scholar] [Publisher Link]

[6] Najim Dehak et al., "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[7] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "VoxCeleb: A Large-scale Speaker Identification Dataset," *arXiv preprint*, pp. 1-6, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman, "VoxCeleb2: Deep Speaker Recognition," *arXiv preprint*, pp. 1-6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Muhammad Mohsin Kabir et al., "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," *IEEE Access*, vol. 9, pp. 79236-79263, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Dávid Sztahó, György Szaszák, and András Beke, "Deep Learning Methods in Speaker Recognition: A Review," *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 65, no. 4, pp. 310-328, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Wenzao Li et al., "TDNN Achitecture with Efficient Channel Attention and Improved Residual Blocks for Accurate Speaker Recognition," *Scientific Report*, vol. 15, pp. 1-13, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[12] David Snyder et al., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, Canada, pp. 5329-5333, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[13] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," *Inter Speech*, Shanghai, China, pp. 3830-3834, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Kaiming He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[15] Alexei Baevski et al., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *NeurIPS Proceedings: 34th Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 1-12, 2020. [Google Scholar] [Publisher Link]

[16] Wei-Ning Hsu et al., "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451-3460, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Tessfu Geteye Fantaye, Junqing Yu, and Tulu Tilahun Hailu, "Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition," *Computers*, vol. 9, no. 2, pp. 1-27, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[18] Jaher Hassan Chowdhury, Sheela Ramanna, and Ketan Kotecha, "Speech Emotion Recognition with Light Weight Deep Neural Ensemble Model using Hand Crafted Features," *Scientific Report*, vol. 15, pp. 1-14, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] Zhongxin Bai, and Xiao-Lei Zhang, "Speaker Recognition based on Deep Learning: An Overview," *Neural Networks*, vol. 140, 2021, pp. 65-99, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Daniele Salvati, Carlo Drioli, and Gian Luca Foresti, "A Late Fusion Deep Neural Network for Robust Speaker Identification using Raw Waveforms and Gammatone Cepstral Coefficients," *Expert Systems with Applications*, vol. 222, pp. 1-9, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yi Liu et al., "Introducing Phonetic Information to Speaker Embedding for Speaker Verification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[22] Soufiane Hourri, and Jamal Kharroubi, "A Deep Learning Approach for Speaker Recognition," *International Journal of Speech Technology*, vol. 23, pp. 123-131, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[23] Wu Zunjing, and Cao Zhigang, "Improved MFCC-Based Feature for Robust Speaker Identification," *Tsinghua Science and Technology*, vol. 10, no. 2, pp. 158-161, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[24] V. Sabitha, and P. Janardhanan, "Performance Analysis of Speaker Identification System using MFCC and DWT under Various Noise Levels," *International Journal of Engineering Research & Technology*, vol. 2, no. 6, pp. 3337-3341, 2013. [Google Scholar] [Publisher Link]

[25] Haris Isyanto, Ajib Setyo Arifin, and Muhammad Suryanegara, "Voice Biometrics for Indonesian Language Users using Algorithm of Deep Learning CNN Residual and Hybrid of DWT-MFCC Extraction Features," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 13, no. 5, pp. 622-634, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[26] Amit Moondra, and Poonam Chahal, "Speaker Recognition Improvement for Degraded Human Voice using Modified-MFCC with GMM," *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 14, no. 6, pp. 246-252, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[27] Zhiyi Ji et al., "Speaker Recognition System based on MFCC Feature Extraction CNN Architecture," *Academic Journal of Computing & Information Science*, vol. 7, no. 7, pp. 47-59, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[28] Manish Tiwari, and Deepak Kumar Verma, "Enhanced Text-independent Speaker Recognition using MFCC, Bi-LSTM, and CNN-based Noise Removal Techniques," *International Journal of Speech Technology*, vol. 27, pp. 1013-1026, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[29] Nourah M. Almarshady, Adal A. Alashban, and Yousef A. Alotaibi, "Analysis and Investigation of Speaker Identification Problems Using Deep Learning Networks and the YOHO English Speech Dataset," *Applied Sciences*, vol. 13, no. 17, pp. 1-19, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[30] Giovanni Costantini, Valerio Cesarini, and Emanuele Brenna, "High-Level CNN and Machine Learning Methods for Speaker Recognition," *Sensors*, vol. 23, no. 7, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[31] Sonia Malik, Deepti Deshwal, and Neelu Trivedi, "AI-Driven Speaker Identification: Enabling Human-Centric and Trustworthy Intelligent Systems," *Human-Centric Intelligent Systems*, vol. 6, pp. 141-162, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] Pinyan Li et al., "Enhancing Speaker Recognition with CRET Model: a fusion of CONV2D, RESNET and ECAPA-TDNN," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2025, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[33] Zahra Shah, Giljin Jang, and Adil Farooq, "Feature Fusion for Performance Enhancement of Text Independent Speaker Identification," *ICCK Transactions on Intelligent Systematics*, vol. 2, no. 1, pp. 27-37, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[34] Neha Chauhan, Tsuyoshi Isshiki, and Dongju Li, "Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies," *Acoustics*, vol. 6, no. 2, pp. 439-469, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Fereshteh Manafzadeh Heir, Hossein Najafzadeh, and Sarvenaz Erfani, "A Hybrid CNN and Reinforcement Learning Framework for Speaker Identification using Mel-Spectrogram and Continuous Wavelet Transform Features," *Scientific Reports*, vol. 16, pp. 1-27, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[36] Shalini Tomar, and Shashidhar G. Koolagudi, "CNN-MFCC Model for Speaker Recognition using Emotive Speech," *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] P. Sandhya et al., "Spectral Features for Emotional Speaker Recognition," *2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, Bengaluru, India, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[38] Shaik Riyaz, Bathula Lakshmi Bhavani, and S. Venkatrama Phani Kumar, "Automatic Speaker Recognition System in Urdu using MFCC & HMM," *International Journal of Recent Technology and Engineering*, vol. 7, no. 5S4, pp. 109-113, 2019. [Google Scholar] [Publisher Link]

[39] Havva Çelıktaş, and Cemal Hanılçı, "A Study on Turkish Text — Dependent Speaker Recognition," *2017 25th Signal Processing and Communications Applications Conference (SIU)*, Antalya, Turkey, pp. 1-4, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[40] Mansour Alsulaiman, Awais Mahmood, and Ghulam Muhammad, "Speaker Recognition based on Arabic Phonemes," *Speech Communication*, vol. 86, pp. 42-51, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[41] S. Davis, and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980. [CrossRef] [Google Scholar] [Publisher Link]

[42] Meinard Mülle, *Fundamentals of Music Processing*, 1st ed., Springer, pp. 1-487, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[43] M. R. H. Mondal, Subrato Bharati, and Prajoy Podder, "Diagnosis of COVID-19 Using Machine Learning and Deep Learning: A Review," *Current Medical Imaging*, vol. 17, no. 12, pp. 1403-1418, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Zohaib Mushtaq, Shun-Feng Su, and Quoc-Viet Tran, "Spectral Images based Environmental Sound Classification using CNN with Meaningful data Augmentation," *Applied Acoustics*, vol. 172, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[45] Meysam Vakili, Mohammad Ghamsari, and Masoumeh Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification," *arXiv preprint*, pp. 1-13, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[46] Serkan Keser, and Esra Gezer, "Comparative Analysis of Speaker Identification Performance using Deep Learning, Machine Learning, and Novel Subspace Classifiers with Multiple Feature Extraction Techniques," *Digital Signal Processing*, vol. 156, 2025, [CrossRef] [Google Scholar] [Publisher Link]

[47] D.A. Reynolds, and R.C. Rose, "Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[48] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462, 1997. [CrossRef] [Google Scholar] [Publisher Link]

[49] Björn Schuller et al., "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication*, vol. 53, no. 9-10, pp. 1062-1087, 2011. [CrossRef] [Google Scholar] [Publisher Link]