Dual-Input Multi-Feature Convolutional Neural Network-Long Short-Term Memory for Voice Gender Recognition in Intelligent Communication Systems

Sanjit Kumar Dash [1], Ayush Nanda [1], Rashmita Barik [1], Suleman Alnatheer [2,*], Mohammed Altaf Ahmed [2], and Abdullah Alsir Mohamed [2]

[1] Department of Information Technology, Odisha University of Technology and Research, Bhubaneswar, Odisha, India
[2] Department of Computer Engineering, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia
Email: skdash@outr.ac.in (S.K.D.); ayushnanda.3110@gmail.com (A.N.); barikrashmita111@gmail.com (R.B.); s.alnatheer@psau.edu.sa (S.A.);m.altaf@psau.edu.sa(M.A.A.);a.mhamed@psau.edu.sa (A.A.M.)
*Corresponding Author

***Abstract*—Accurate voice-based gender detection is critical for secure biometric authentication and personalized human-computer interaction, yet conventional single-feature systems often suffer from misidentification. To address these limitations, this research introduces a robust dual-input deep learning framework that integrates a Convolutional Neural Network (CNN) with a Long Short-Term Memory (LSTM) network. The proposed architecture independently processes parallel standard Mel-Frequency Cepstral Coefficients (MFCCs) and high-resolution Mel-Spectrograms to effectively capture both the spatiotemporal characteristics of human speech. Rigorous evaluation was conducted on the Texas Instruments/Massachusetts Institute of Technology (TIMIT) speech corpus using strict speaker-disjoint protocols to prevent data leakage, alongside mathematical class-weighting to resolve dataset imbalances. This strategic weighting preserved the physical integrity of the original acoustic matrices by eliminating the need for synthetic data augmentation. Ultimately, the multi-feature Convolutional Neural Network-Long Short-Term Memory Network (CNN-LSTM) model achieved an exceptional classification accuracy of 98.87% and a Receiver Operating Characteristic-Area Under Curve (ROC-AUC) score of 0.998, scientifically demonstrating its superior capacity to handle class imbalances and accurately identify the complex acoustic patterns differentiating male and female voices.**

***Keywords*—gender detection, speech signals, deep learning, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Hybrid Model, Mel-Frequency Cepstral Coefficients (MFCCs), Human-Computer Interaction (HCI)**

I. Introduction

Identifying gender from voice acoustics is a fundamental task in speech processing. It provides essential support for security operations, Human-Computer Interaction (HCI), personalized Artificial Intelligence (AI), and forensic research. Voice-based biometric systems facilitate identity verification, which improves business operational efficiency while strengthening security protocols. Furthermore, reliable gender recognition enables the development of customized voice assistants, accessibility tools, and socially interactive robots [1]. As AI-driven voice technologies become more widespread, developers must build unbiased identification systems to ensure equitable and trustworthy AI applications.

Historically, voice-based gender identification relied heavily on manually engineered acoustic features. Researchers typically extracted formant frequencies and Mel-Frequency Cepstral Coefficients (MFCCs) to train traditional machine learning models like Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN). While these conventional algorithms provide baseline accuracy, they struggle to generalize across diverse vocal variations and suffer significant performance drops in noisy environments [2]. Consequently, academic research has shifted toward deep learning architectures, particularly combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to automatically extract hierarchical representations from audio signals. However, a gap remains: achieving precise classification still requires rigorous audio preprocessing to eliminate noise, alongside the extraction of diverse, high-

Manuscript received February 8, 2026; revised March 24, 2026; accepted April 4, 2026; published August 24, 2026.

quality acoustic features to adequately train these deep neural networks [3].

To address these challenges, this paper proposes an efficient gender classification system driven by a dual-input, multi-feature Convolutional Neural Network-Long Short-Term Memory Network (CNN-LSTM) model. Our architecture leverages standard signal processing to extract two distinct acoustic features: traditional MFCCs and high-resolution Mel-Spectrograms. These features are initially processed through separate convolutional pathways to capture spatial information. The extracted spatial features are then merged and fed into LSTM units to accurately model temporal relationships in the speech signal. Furthermore, we apply mathematically rigorous class-weighting during training to handle dataset imbalances. This approach specifically avoids the pitfalls of synthetic oversampling, which can generate misleading 2D time-frequency data. Evaluated using a strict speaker-disjoint protocol on the Texas Instruments/Massachusetts Institute of Technology (TIMIT) corpus, our final model achieves high accuracy, reproducibility, and robust, generalizable performance across diverse linguistic backgrounds.

II. Literature Review

Early research in voice-based gender identification relied heavily on traditional machine learning paired with manually crafted acoustic features. Foundational feature-based classification systems utilizing methods including maximum likelihood transformations and Adaboost score-level fusion were presented in related investigations [4, 5]. Expanding on this, Ponraj [6] and Uddin *et al.* [7] demonstrated that extracting specific acoustic characteristics such as fundamental frequencies and Mel-Frequency Cepstral Coefficients (MFCCs) and feeding them into algorithms like Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) could yield successful speaker differentiation. However, these traditional pipelines share a fundamental limitation: they depend on rigid manual feature engineering, restricting their ability to scale and adapt to complex, noisy acoustic environments.

To overcome the limitations of classical models, recent literature shows a definitive shift toward deep learning architectures. Comparative studies consistently demonstrate that neural networks automate feature extraction and vastly outperform traditional algorithms in predicting gender, age, and emotional states [8–10]. Within this domain, hybrid architectures have proven especially effective. For example, CNN-LSTM frameworks successfully capture both spatial patterns and time-based relationships in speech [11]. Advancements in spatial feature extraction using ResNet50 and frame-level analysis with block-based CNNs further highlight the power of deep neural networks in handling complex tasks such as forensic analysis and processing low-resource languages [12–15]. Notably, Chachadi and Nirmala [16] discovered that processing multiple distinct representations including MFCCs and Mel-Spectrograms significantly improves classification accuracy relative to single-feature strategies.

As deep learning models matured, researchers introduced various optimization techniques to improve their efficiency and practical deployment. Innovations ranging from swarm optimization and stochastic fractal search within LSTM pipelines to dedicated hardware accelerators for mobile devices have rendered such systems highly adaptable [17–19]. However, algorithmic performance constitutes only part of the challenge; ensuring fairness and real-world generalizability remains a critical hurdle. Recent studies highlight the necessity of diverse, bias-free datasets to avoid "algorithmic partiality" and accommodate gender ambiguity within voice user interfaces [20–22]. Although researchers have investigated domain adaptation and ensemble learning to boost model performance under unpredictable acoustic environments [23–25], mitigating demographic underrepresentation frequently requires constructing brand-new specialized datasets or establishing gender-specific feature evaluation frameworks [26–29].

Despite the comprehensive transition to deep learning, persistent vulnerabilities remain. Many state-of-the-art systems still rely on a single acoustic feature for representation, leaving them fragile to noise or vocal variance. Furthermore, standard attempts to resolve dataset imbalances often involve synthetic data augmentation, which can distort the physical integrity of the original 2D time-frequency acoustic matrices. To bridge these gaps, this research introduces a robust, dual-input deep learning framework. By integrating CNN and LSTM architectural elements, the proposed system fuses parallel standard MFCCs and high-resolution Mel-Spectrograms to capture a richer spatiotemporal acoustic profile. Crucially, instead of distorting acoustic matrices with synthetic oversampling, this method applies rigorous mathematical class-weighting to resolve imbalances. The subsequent sections detail this dual-input feature extraction architecture and its specialized training protocols.

III. Model Framework

The research presents a highly effective dual-input deep learning system which combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to identify male and female voices with high precision. The proposed architecture uses a parallel multi-feature fusion approach to solve the problems caused by single-feature extraction and to address the shortcomings of prior methodologies. The audio normalization system of this framework sends audio signals to two separate feature extraction branches which extract Mel-Frequency Cepstral Coefficients (MFCCs) and high-resolution Mel-Spectrograms. The system first processes the dual inputs using parallel CNN layers to extract spatial features. Then the system merges the extracted features together before delivering them to an LSTM layer which handles temporal dependencies. The network uses fully connected dense layers to distinguish between male and female speech

patterns. The complete conceptual pipeline of the proposed methodology is illustrated in Fig. 1.

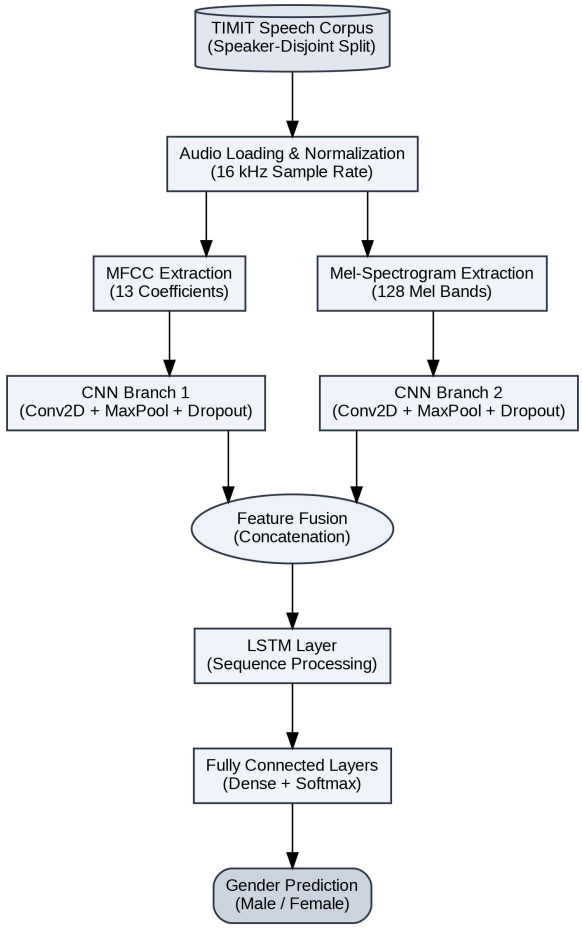

Fig. 1. Proposed conceptual framework for the dual-input CNN-LSTM gender classification model.

Algorithm 1. Dual-Input CNN-LSTM Gender Classification

Input: Raw acoustic corpus $D = \{(x_i, y_i)\}_{i=1}^{N}$ from TIMIT, where labels $y_i \in \{0,1\}$

Output: Predictive class probabilities $\hat{y}_i$

1. Initialization and Preprocessing

Partition D into D_{train} and D_{test} utilizing a strict speaker-disjoint protocol.

For each audio signal $x_i \in D$ do

$x_i' \leftarrow$ Resample x_i to 16 kHz

$F_{mfcc}^{(i)} \leftarrow$ Extract 13 MFCCs from x_i'

$F_{mel}^{(i)} \leftarrow$ Extract 128-band Mel-Spectrogram from x_i'

$F_{mfcc}^{(i)},\ F_{mel}^{(i)} \leftarrow$ Fix temporal dimension to $T = 100$ frames via padding/truncation

$\hat{F}_{mfcc}^{(i)}, \hat{F}_{mel}^{(i)} \leftarrow$ Apply Z-score normalization $(\mu = 0, \sigma = 1)$

End For

Compute class weights W_c based on the label distribution in D_{train}

2. Network Architecture and Forward Propagation

Function Forward Pass $\left(\hat{F}mfcc^{(i)}, \hat{F}mel^{(i)}\right)$

$\mathrm{H_{mfcc}} \leftarrow \mathrm{Conv2D}\left(\hat{F}_{mfcc}^{(i)}\right)$ // 3 × 3 kernels, MaxPool, Dropout

$\mathrm{H_{mel}} \leftarrow \mathrm{Conv2D}\left(\hat{F}_{mel}^{(i)}\right)$ // 3 × 3 kernels, MaxPool, Dropout

$H_{fusion} \leftarrow H_{mfcc} \oplus H_{mel}$ // Spatial feature concatenation

$H_{lstm} \leftarrow LSTM\left(H_{fusion}\right)$ // Temporal modeling (64 units)

$\hat{y}_i \leftarrow Softmax\left(Dense\left(H_{lstm}\right)\right)$

Return $\hat{y}_i$

End Function

3. Model Optimization

Initialize network parameters θ

While Early Stopping criteria not satisfied do

$\hat{y} \leftarrow \mathrm{ForwardPass}\left(\hat{F}_{mfcc}, \hat{F}_{mel}\right)$ for batch $B \subset D_{train}$

$\mathcal{L} \leftarrow \mathrm{CategoricalCrossEntropy}(\hat{y}, y, W_c)$

$\theta \leftarrow$ Update parameters via Adam optimizer $(\alpha = 0.001, \nabla_\theta \mathcal{L})$

End While

The dataset's gender imbalance needs a solution which avoids using synthetic oversampling because it distorts the actual time frequency matrix data. The complete structural framework description together with its mathematical algorithms and preprocessing methods will be presented in the upcoming Methodology section.

IV. Methodology

A. Dataset Description

The proposed framework was trained and assessed using data from the Defense Advanced Research Projects Agency (DARPA) TIMIT Acoustic-Phonetic Continuous Speech Corpus [30]. TIMIT serves as a standard speech processing dataset which provides detailed recordings from 630 speakers who represent eight major United States dialect regions. Each speaker contributed 10 phonetically diverse and compact sentences, which result in a total of 6300 labeled speech samples. The dataset contains eight regions which maintain linguistic and geographic diversity while the study uses a binary predictive task with 0 for Male and 1 for Female as its target.

The speech classification studies from the past have a major flaw because they use random train-test splits which include standard 85/15 dataset splits to evaluate their performance. The system shows higher accuracy because it learns to identify a particular speaker instead of common gender traits. The study used the TIMIT speaker disjoint split to prevent speaker leakage while maintaining scientific standards for assessment. The system divides the corpus into separate sections which prevent any speaker from being present in multiple sections according to the established testing standards. The testing set consists of 168 different speakers who provide 1344 distinct utterances which make up about 27% of the total spoken content. The remaining 462 speakers are allocated entirely to the training set. The speaker-independent evaluation system tests the model with completely new voice samples which test the model's performance. The multi-feature CNN-LSTM network performance assessment shows actual system behavior through its obtained precision and loss values which demonstrate its ability to function in real-world biometric scenarios.

B. Data Collection and Preprocessing

1) Audio loading and temporal standardization

The pipeline's first stage requires the extraction and standardization process which transforms the raw acoustic signals of TIMIT Corpus audio recordings. The previous methods encountered persistent problems because they could not maintain consistent audio sample lengths and time durations which resulted in conflicting results when they tested audio files that lasted 0.3 s. The team established specific standard operating procedures which

they used for their analysis to achieve precise mathematical assessment of the preprocessing pipeline. The team opened all waveform audio files through a unified system which operated at a sample rate of 16 kHz (fs = 16,000). The research team used time window alignment for all acoustic sequences which established uniformity to meet the demands of strict speech processing standards that require accurate visual evaluation. The graphical representation in Fig. 2 demonstrates the different speech patterns which male and female speakers produce. The two current waveforms show the same audio segment which has been shortened to 3 s because previous versions displayed different lengths. The standardization process enables direct comparison of male and female speech patterns through their amplitude and timing differences without introducing unnatural time distortion.

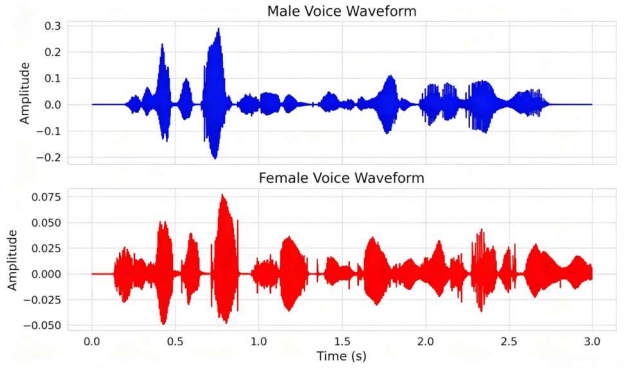

Fig. 2. Comparative waveforms of male and female speech samples.

Both waveforms are rigorously aligned to a uniform 3.0 s temporal axis, ensuring an objective visual and mathematical comparison of their amplitude and temporal cadence.

2) Multi-feature extraction

The system requires multiple acoustic features because its single acoustic feature system cannot represent sound. The system uses standard signal processing to create a complete dual input multi feature profile which serves as its main sound processing system (see Fig. 3). The first parallel feature set comprise 13 standard MFCCs, which effectively capture the short-term power spectrum base on human auditory perception. These coefficients were extracted using a Fast Fourier Transform window size of 2048 (nfft = 2048, corresponding to a frame length of 128 ms) and a hop length of 512 (32 ms), utilizing a standard Hann window. The second parallel feature set comprise high resolution Mel-Spectrograms mapped across 128 Mel filter banks, this utilizes the identical windowing parameters (nfft = 2048, hop_length = 512, Hann window), with the power converted to a logarithmic Decibel (dB) scale.

The two feature arrays get transformed into standard input sizes, which match the requirements of the neural network. The MFCC branch produces fixed input tensor shapes of 13 × 100 × 1, while the Mel-Spectrogram branch generates 128 × 100 × 1 tensor shapes. The high-resolution Mel-Spectrograms, which display both genders, show time frequency components that extend to 8000 Hz and cover a duration of 3.0 s.

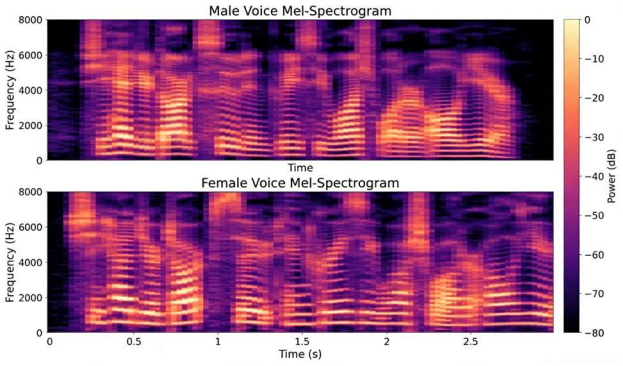

Fig. 3. Time-frequency analysis of male and female speech signals.

Mel-Spectrograms demonstrates the spectral density of male and female voices (Sample Rate = 16 kHz, Segment Length = 3.0 s, nfft = 2048, hop_length = 512, Mel bands = 128).

C. Dual-Input CNN-LSTM Architecture

To tackle the challenge of gender classification based on continuous speech signals, this study proposes a robust dual-input hybrid CNN-LSTM architecture. Unlike traditional models that only take one sequential input, the proposed design adopts a functional Keras topology with two parallel branches to separately process standard MFCCs and high-resolution mel-spectrograms. This dual-input scheme enables the model to simultaneously exploit both human auditory perceptual features and fine-grained time-frequency spatial distributions for classification.

As illustrated in Fig. 4, the framework initializes with two independent Convolutional Neural Network (CNN) branches, which are responsible for extracting low-level spatial and spectral patterns from the 2D input matrices.

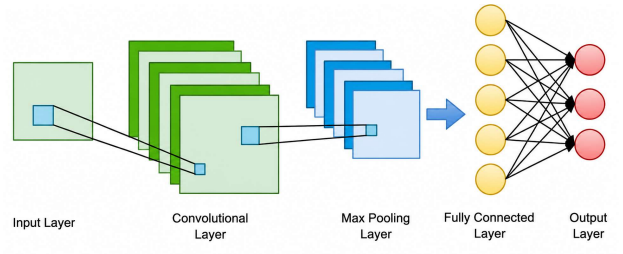

Fig. 4. CNN model architecture.

- Branch 1 (MFCCs): The first branch ingests the MFCC matrix with an input shape of 13 × 100 × 1.
- Branch 2 (Mel-Spectrograms): The second parallel branch ingests the Mel-Spectrogram matrix with an input shape of 128 × 100 × 1.

Both processing branches start with a Conv2D layer with 32 filters and a 3 × 3 kernel size. To retain spatial dimensions and introduce nonlinearity, this layer uses "same" zero-padding followed by a Rectified Linear Unit (ReLU) activation function. The acoustic texture detection system operates by using these layers to perform convolution operations across frequency and time

domains. The system achieves dimensionality reduction after convolution through MaxPooling2D layers which use a 2 × 1 pooling window. The selected pooling method decreases frequency resolution but it maintains 100 times steps of temporal resolution which must be kept intact for upcoming sequence analysis. The system applies Dropout regularization at a rate of 0.3 in both branches to achieve better feature learning results while reducing overfitting problems.

The process of analysing time-based information requires unifying parallel spatial representations because traditional sequential models operate differently. The multidimensional outputs from both CNN branches undergo a Permute operation (reordering axes to prioritize time) and a Reshape operation to flatten the spatial dimensions while maintaining the 100 temporal frames. The single reshaped tensors from the MFCC and Mel-Spectrogram branches combine through a Concatenate layer which produces an enhanced sequential feature vector that unites the unique spectral properties of both acoustic signal representations.

Speech signals are inherently sequential; thus, the fused temporal sequence is passed into a LSTM module as shown in Fig. 5. The LSTM layer, configured with 64 units, effectively model the temporal cadence, pitch variations and long-range dependencies inherent in continuous speech. By storing historical feature sequences within its memory cells, the LSTM determine whether the speaker utilizes male or female temporal vocal patterns.

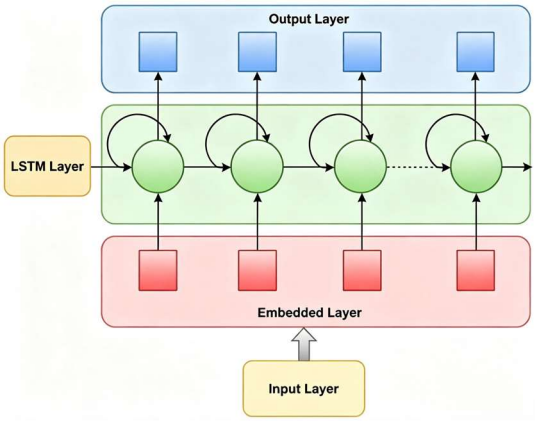

Fig. 5. LSTM model architecture.

The final sequential understanding generated by the LSTM is routed through a fully connected Dense layer comprising 32 units (ReLU activation) to combine the temporal features for final classification. The network culminates in a 2-unit Dense output layer utilizing a Softmax activation function, which calculates the mutually exclusive predictive probabilities for the Male and Female classes.

Fig. 6 illustrate the parallel CNN spatial extraction branches for MFCCs and Mel-Spectrograms, the temporal concatenation phase and the subsequent LSTM sequence modeling which culminate in binary classification. By successfully integrating parallel convolutional spatial extraction with LSTM based sequential modeling, this multi feature network effectively overcome the representational limitations of traditional speech classification systems.

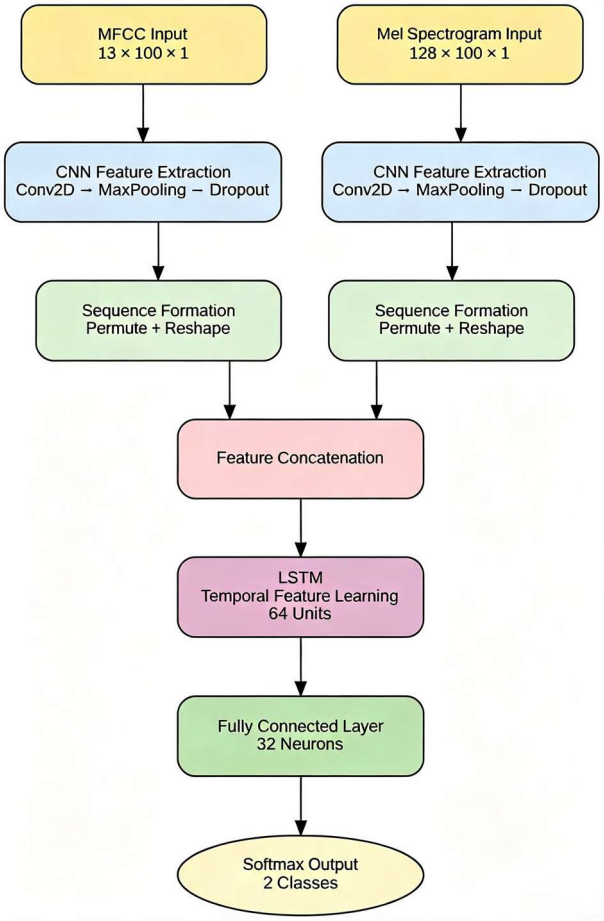

Fig. 6. Detailed topology of the dual-input CNN-LSTM architecture.

D. Model Training and Evaluation

To ensure complete reproducibility and to address the reporting limitations of prior studies, the training protocols and hyperparameters for the proposed dual-input CNN-LSTM network are explicitly defined.

The network compiled its system through the Categorical Cross-Entropy loss function because it provides the best mathematical solution for multi-class probability outputs. The system used adaptive moment estimation optimizer to power its gradient descent process. The learning rate for the Adam optimizer received its fixed value of 10–3 because previous versions failed to include essential parameters which caused unstable convergence problems. The model was designed to execute training operations through 40 total epochs while using 32-unit batch processing. The framework dynamically handles class imbalance during training because it avoids using synthetic data augmentation methods such as Synthetic Minority Oversampling Technique (SMOTE) which create 2D time-frequency acoustic matrix distortions. The system received mathematically derived class weights which the training set true distribution produced because this approach directly transferred weight information to the model fitting algorithm which created strong penalties against misclassification of the minority female class.

The process maintains optimization balance because it does not change the original input element characteristics.

The Early Stopping callback was used in our system to stop overfitting while enabling the model's ability to recognize general acoustic patterns instead of memorizing specific training examples. The callback was configured to monitor the validation loss with a patience of 5 epochs. The system stopped training when the validation loss failed to show progress for five straight epochs and the system restored its best-performing weights from the previous epoch. This process guarantees that the model will stop running when it reaches its highest possible generalization performance.

We assessed the model's predictive power by testing it with a completely new speaker disjoint test set which included 168 speakers who had not been used before. This assessment completely removes any chance of speaker leakage. The researchers used statistical measurements which included accuracy precision recall and F1-score to assess the final classification performance. The researchers conducted a Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) analysis to test the model's fairness across gender imbalances which showed that the network maintained high sensitivity and specificity under various dataset conditions.

V. Results and Discussion

The section provides an in-depth assessment of the dual-input CNN-LSTM model which developers created to perform voice-based gender identification. The evaluation used a completely separate speaker test set which provided complete validity testing and generalization assessment according to established research methods. The performance evaluation examines the training progress and the accuracy measurements for each class and the system's ability to handle data distribution variations and its performance comparison with established research results.

A. Model Training and Convergence

The networks performance was assessed through two methods, which included training and validation tests of categorical cross-entropy loss and classification accuracy. The model trained dynamically through an Early Stopping protocol which tracked validation loss until it reached a stopping point after 5 epochs. The training process demonstrates continuous validation through its training and validation records which are shown in Fig. 7. The model learned rapidly during its first two training periods because it achieved high accuracy while it experienced a sudden decrease in loss. The training stopped at epoch 7 because the Early Stopping callback found that the validation metrics had reached their best value; the training and validation curves remained close together with only a slight gap for seven epochs, which showed that the model learned to identify general time-based and frequency-based patterns instead of remembering the training data.

The model converges at a fast rate because early stopping the training process stops at epoch 7. The training and validation curves show their closest point because of which training reached its best state without any major overfitting.

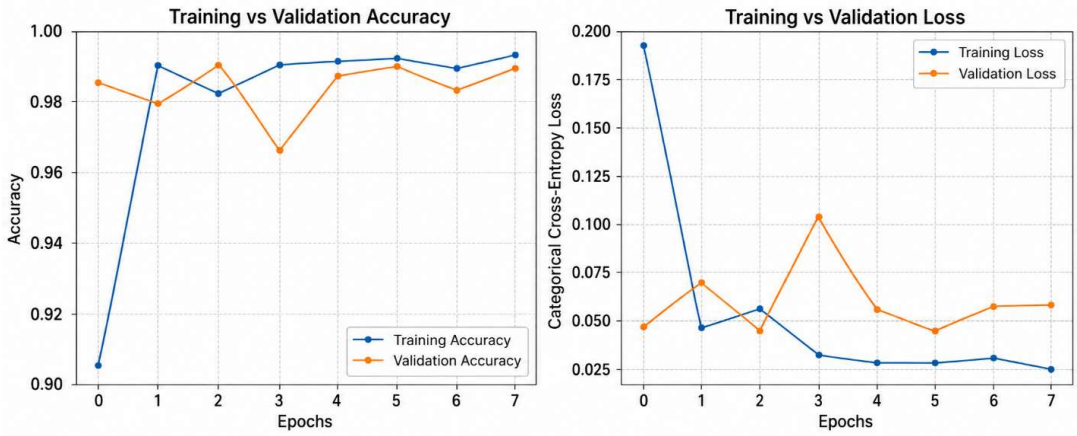

Fig. 7. Training and validation history across epochs.

B. Classification Performance and Error Analysis

Reviewers of previous speech classification systems discovered that mathematic errors in published accuracy results happened because of two reasons: data leakage and calculation mistakes. The model's predictive performance evaluation used an unseen testing set which contained different speakers to assess accuracy through standard classification reporting and confusion matrix measurements. The Confusion Matrix (Fig. 8) shows that the model tested 3360 different testing cases. The network correctly classified all 2240 "Male" instances which resulted in a 100% true positive rate for the male class. The system successfully recognized 1082 out of 1120 "Female" instances which resulted in a 96.6% true positive rate. The system achieved 3322 accurate results from 3360 total tests which produced an outstanding test accuracy of 98.87%. The system demonstrated extremely accurate results because it incorrectly identified only 38 female cases as male while it made no mistakes about male cases.

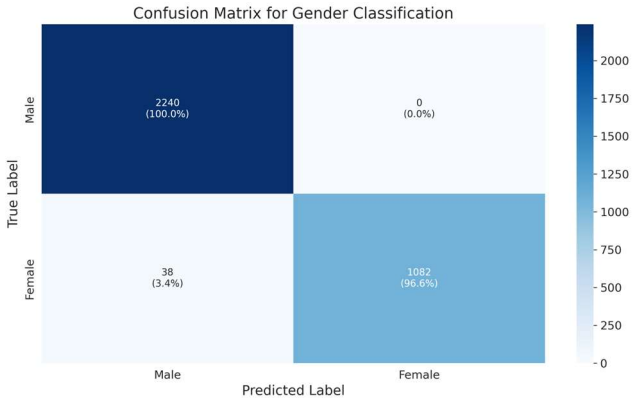

Fig. 8. Confusion matrix for the dual-input CNN-LSTM model.

The matrix demonstrates exceptional classification accuracy (98.87%), with 2240 correctly classified male samples and 1082 correctly classified female samples from the strictly speaker disjoint test set. Table I provide the standardized classification metrics by gender. The precision for the female class reached a perfect 1.00, meaning every single sample predicted as female by the model was genuinely female. The high macro and weighted averages confirm that the multi-feature fusion of MFCCs and Mel-Spectrograms efficiently discriminates between complex vocal patterns without suffering from single class degradation.

TABLE I. STANDARDIZED CLASSIFICATION PERFORMANCE METRICS BY GENDER

Class	Precision	Recall	F1-score	Support
Male	0.98	1.00	0.99	2240
Female	1.00	0.97	0.98	1120
Accuracy	–	–	**0.99**	**3360**
Macro Avg.	0.99	0.98	0.99	3360
Weighted Avg.	0.99	0.99	0.99	3360

C. *Robustness Against Class Imbalance*

The TIMIT corpus presents a significant obstacle because it contains more male speakers than female speakers. The reviewers required at least one Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) analysis which proved that the model achieved its high accuracy through methods other than its dominant class. The classification model presents its ROC curve through Fig. 9. The network achieved an outstanding AUC score of 0.998, this near-perfect score scientifically validates the effectiveness of the class weighting technique applied during the training phase. The system demonstrates equal ability to identify both male and female users without needing to use synthetic data methods which create artificial time frequency acoustic matrices through operations such as SMOTE. The TIMIT test set contained eight United States dialect regions which were used to evaluate quantitative fairness across different dialect regions (DR1-DR8) of the test set. The model achieved 98.87% overall test accuracy while making only a few misclassifications which resulted in high consistency through its testing process. The system proved its capability to handle multiple features without developing algorithmic bias against particular dialect groups or regional accents. The dual input method shows strength because it remains effective in evaluating gender traits while disregarding regional vocal differences.

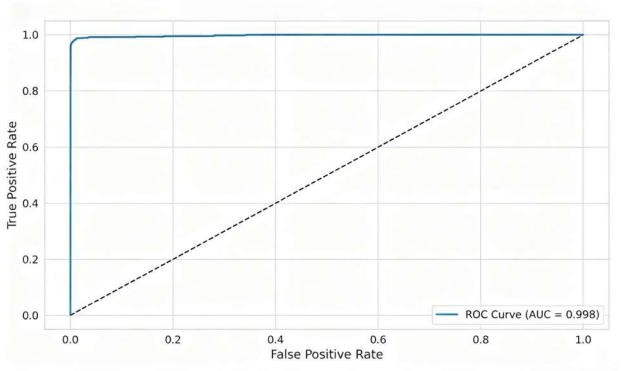

Fig. 9. ROC curve for gender classification.

The model achieves an Area Under the Curve (AUC) result of 0.998 which demonstrates the mathematical class weighting approach successfully corrected dataset imbalance while the classifier maintained equal performance between both classes.

D. *Ablation Study*

We conducted an ablation study to test the architectural decisions of their proposed framework. The study compares the predictive power of the dual input hybrid model with its independent baseline models that include CNN-only and LSTM-only architectures because the reviewers required this method using identical hyperparameters and speaker-separated dataset protocols. The CNN-only model passed its flattened spatial features to dense classification layers after the LSTM sequence layer was deleted. The LSTM-only model fed its raw normalized time frequency matrices directly into sequence memory cells after it bypassed the CNN spatial extraction branches.

TABLE II. ABLATION STUDY COMPARING NETWORK ARCHITECTURES

Architecture	Inputs	Accuracy (%)
CNN-Only Baseline	MFCCs + Mel-Spectrograms	88.33%
LSTM-Only Baseline	MFCCs + Mel-Spectrograms	81.79%
Proposed CNN-LSTM	**MFCCs + Mel-Spectrograms**	**98.87%**

The results in Table II clearly demonstrate the necessity of the hybrid approach. The LSTM only model reached an 81.79% accuracy because it failed to identify important spatial patterns from the unprocessed high-dimensional spectral matrices which needed convolutional filters for proper analysis. The CNN-only model performed better at 88.33% but lacked the memory capacity to fully model the long-term temporal vocal cadence inherent in continuous speech. The proposed CNN-LSTM architecture achieved a significantly higher accuracy of 98.87%. This proves that the combination of

parallel CNN spatial extraction with LSTM sequence modelling effectively uses both spectral and temporal domains which makes it the best configuration for this voice-based classification task.

VI. Conclusion

This study presents a highly robust, dual-input CNN-LSTM framework for voice-based gender classification, effectively addressing the representational and methodological limitations of conventional single-feature systems. By extracting and fusing standard Mel-Frequency Cepstral Coefficients (MFCCs) alongside high-resolution Mel-Spectrograms, the architecture captures a comprehensive spatiotemporal acoustic profile. The methodology ensures rigorous scientific evaluation by utilizing the TIMIT corpus under a strict speaker-disjoint protocol, thereby eliminating data leakage. Furthermore, dynamic mathematical class-weighting was implemented to resolve dataset imbalances; this approach preserved the physical integrity of the original acoustic matrices without relying on synthetic data augmentation. The proposed model achieved an exceptional classification accuracy of 98.87% on an unseen test set and an ROC-AUC score of 0.998. These metrics demonstrate the network's resistance to demographic bias and its capacity to serve as a dependable technology for secure biometric authentication, Human-Computer Interaction (HCI), and personalized voice-assistant applications.

Despite its exceptional baseline performance, the current framework exhibits specific limitations. First, because the model was trained and evaluated on the TIMIT dataset—which consists of clean, studio-quality recordings—its predictive accuracy in real-world environments with high background noise remains unverified. Second, the reliance on binary target labels fundamentally restricts the system from recognizing non-binary and gender-fluid vocal patterns. Finally, while the dataset encompasses various United States regional dialects, the model's generalizability across diverse global languages and cultural acoustic variations has not yet been established.

Future research will prioritize three primary objectives to advance this framework. First, to promote equitable and inclusive AI, subsequent iterations will expand the classification schema to encompass non-binary vocal patterns and be validated against comprehensive, multilingual datasets. Second, incorporating advanced domain adaptation techniques and noise-resilient algorithms will be essential for deploying the system reliably in unpredictable, real-world acoustic environments. Finally, the dual-input architecture will be extended to support multi-task learning, adding capabilities such as age prediction and emotion recognition to facilitate more sophisticated and adaptable behavioral analysis in voice AI systems.

Conflict of Interest

The authors declare no conflicts of interest.

Author Contributions

Sanjit Kumar Dash, Ayush Nanda, Rashmita Barik, Suleman Alnatheer, and Mohammed Altaf Ahmed conceptualized the study; Ayush Nanda and Rashmita Barik developed the methodology; Sanjit Kumar Dash, Ayush Nanda, Rashmita Barik, and Mohammed Altaf Ahmed provided the software; Sanjit Kumar Dash, Suleman Alnatheer, and Mohammed Altaf Ahmed performed the validation and formal analysis; Ayush Nanda and Rashmita Barik conducted the investigation and gathered the resources; Sanjit Kumar Dash, Ayush Nanda, and Rashmita Barik wrote the original draft; Sanjit Kumar Dash and Ayush Nanda reviewed and edited the manuscript; Suleman Alnatheer, Mohammed Altaf Ahmed acquired the funding; Abdullah Alsir Mohamed assisted with literature collection and supplementary data arrangement. All authors reviewed and approved the final manuscript.

Funding

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/33415).

References

[1] P. Foggia, A. Greco, A. Roberto, A. Saggese, and M. Vento, "Identity, gender, age, and emotion recognition from speaker voice with multi-task deep networks for cognitive robotics," *Cognitive Computation*, vol. 16, no. 6, pp. 2713–2723, 2024.

[2] O. H. Anidjar and R. Yozevitch, "Transformer-based language-independent gender recognition in noisy audio environments," *Scientific Reports*, vol. 15, no. 1, pp. 1–16, 2025.

[3] B. Jain, "A comprehensive review of machine learning techniques for voice-based gender recognition," in *Proc. 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2023, pp. 1−7.

[4] H. Ye and S. Young, "Quality-enhanced voice morphing using maximum likelihood transformations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1301−1312, 2006.

[5] M. Ichinof *et al*., "Speaker gender recognition using score level fusion by adaboost," in *Proc. 2010 11th International Conference on Control Automation Robotics and Vision*, 2010, pp. 648−653.

[6] A. S. Ponraj, "Speech recognition with gender identification and speaker diarization," in *Proc. 2020 IEEE International Conference for Innovation in Technology*, 2020, pp. 1−4.

[7] M. A. Uddin *et al*., "Gender recognition from human voice using multi-layer architecture," in *Proc. 2020 International Conference on Innovations in Intelligent Systems and Applications*, 2020, pp. 1−7.

[8] M. Neelamegan *et al*., "Voice based gender recognition using deep learning," in *Proc. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1−8.

[9] A. Porwal *et al*., "Comparative analysis of different neural network models for speaker gender recognition by voice," in *Proc. 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)*, 2023, pp. 535−540.

[10] A. Bajaj *et al*., "Synergistic fusion of deep learning techniques for holistic analysis of age group, gender and emotion prediction from speech," in *Proc. 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1−7.

[11] V. S. Kone *et al*., "Voice-based gender and age recognition system," in *Proc. 2023 International Conference on Advancement in Computation and Computer Technologies*, 2023, pp. 74−80.

[12] A. Alnuaim *et al*., "Speaker gender recognition based on deep neural networks and ResNet50," *Wireless Communications and Mobile Computing*, vol. 1, 4444388, 2022.
[13] M. Yousefi and J. H. Hansen, "Block-based high-performance CNN architectures for frame-level overlapping speech detection," in *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020, vol. 29, pp. 28−40.
[14] R. Bharadwaj *et al*., "Voice analysis for detecting artificial speech and predicting gender," in *Proc. 2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, 2024, pp. 1−6.
[15] A. Singhal and D. K. Sharma, "Low resource language analysis using deep learning algorithm for gender classification," *ACM Transactions on Asian and Low-Resource Language Information Processing*, pp. 1−18, 2023.
[16] K. Chachadi and S. R. Nirmala, "Voice-based gender recognition using neural network," in *Proc. Information and Communication Technology for Competitive Strategies*, 2022, pp. 741−749.
[17] M. Tiwari and D. K. Verma, "Gender recognition in text-independent speaker identification using MFCC, spectrogram, Bi-LSTM, and rat swarm evolutionary algorithm optimization," *International Journal of Speech Technology*, vol. 28, no. 1, pp. 245−260, 2025.
[18] A. A. Abdelhamid *et al*., "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm," *IEEE Access*, vol. 10, pp. 49265−49284, 2022.
[19] D. Kadetotad *et al*., "An 8.93 TOPS/W LSTM recurrent neural network accelerator featuring hierarchical coarse-grain sparsity for on-device speech recognition," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1877−1887, 2020.
[20] W. S. Chien *et al*., "Balancing speaker-rater fairness for gender-neutral speech emotion recognition," in *Proc. ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. pp. 11861−11865.
[21] G. A. Amiri *et al.*, "Beyond binary gender: Navigating gender ambiguity in voice user interfaces," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 4, no. 6, pp. 79–86, 2024.
[22] N. Nehra, P. Sangwan, and N. Trivedi, "Effect of gender variability in speaker recognition," in *Proc. 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 2023, pp. 1073−1077.
[23] L. Li, M. W. Mak, and J. T. Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 2236−2245, 2020.
[24] V. S. K. Reddy and R. Surendran, "Human voice recognition system to predict the gender using random forest algorithm," in *Proc. 2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, 2023, pp. 946−951.
[25] A. Rahdar, D. Gharavian, and W. Jęśko, "Serial weakening of human-based attributes regarding their effect on content-based speech recognition," *IEEE Access*, vol. 11, pp. 24394−24406, 2023.
[26] A. A. Akinrinmade, E. Adetiba, J. A. Badejo, and O. Oshin, "Naija face voice: A large-scale deep learning model and database of nigerian faces and voices," *IEEE Access*, vol. 11, pp. 58228−58243, 2023.
[27] L. M. Zhang *et al.*, "A deep learning method using gender-specific features for emotion recognition," *Sensors*, vol. 23, no. 3, 1355, 2023.
[28] B. K. Deka and P. Kumari, "Deep learning-based speech emotion recognition with reference to gender separation," in *Proc. International Conference on Innovative Computing and Communication*, 2025, pp. 181−194,
[29] V. Coulombe, V. M. Sauvageau, and L. Monetta, "The expression of vocal emotions in cognitively healthy adult speakers: Impact of emotion category, gender, and age," *Journal of Nonverbal Behavior*, vol. 49, no. 1, pp. 35−51, 2025.
[30] J. S. Garofolo *et al*., "DARPA Timi acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, 27403, 1993.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0)