Speech-Based Gender Recognition Using Conventional MFCC–SVM and Hybrid CNN–LSTM Deep Learning Framework

Digambar B. Gote
ME (E & TC) Student, D. Y. Patil College of Engg. & Technology, Kolhapur
digambar5585@gmail.com

Prof. Dr.T.B.Mohite-patil
D. Y. Patil College of Engg. & Technology Kolhapur,
hodetxtbmp@gmail.com

Abstract Speech-based gender recognition is an important task in intelligent human–computer interaction, biometric authentication, multimedia indexing, and adaptive voice-driven systems. This article presents a complete study of gender recognition from speech by covering both conventional machine learning and deep learning paradigms. Initially, a traditional framework based on Mel-Frequency Cepstral Coefficients (MFCC) and Support Vector Machine (SVM) is implemented and evaluated as a baseline model. Although this approach provides acceptable performance in controlled conditions, its accuracy decreases under noise, channel mismatch, and speaker variability. To overcome these limitations, a hybrid deep learning model combining Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and an attention mechanism is proposed. The CNN extracts robust local spectral patterns from time–frequency speech representations, while the LSTM captures temporal dependencies across successive speech frames. The attention module further improves classification by emphasizing informative voiced regions and suppressing redundant components. Experimental analysis includes feature comparison, confusion matrix analysis, ROC curve evaluation, ablation study, noise robustness, and cross-dataset generalization. Results show that the proposed deep learning model significantly outperforms the conventional MFCC+SVM framework in terms of accuracy, precision, recall, F1-score, and robustness. The proposed system therefore provides an effective and scalable solution for practical speech-based gender recognition.
Keywords: Speech Processing, Gender Recognition, MFCC, CNN–LSTM, Deep Learning

I. Introduction

Speech is one of the most natural and efficient means of human communication, and it carries a large amount of information beyond linguistic content. In addition to spoken words, speech also conveys speaker-specific traits such as gender, age, emotion, accent, and identity. Among these, gender recognition from speech has gained considerable research attention because of its usefulness in speech analytics, personalized user interaction, biometric systems, virtual assistants, call routing, and intelligent multimedia retrieval. An automatic gender recognition system can improve the adaptability of interactive systems and enhance the performance of broader speech processing pipelines.

Traditional approaches to speech-based gender recognition have relied mainly on manually designed acoustic features such as pitch, formants, and Mel-Frequency Cepstral Coefficients (MFCC). These features are then provided to machine learning classifiers such as Support Vector Machines (SVM), decision trees, or k-nearest neighbors. While such methods are computationally simple and easy to interpret, they strongly depend on handcrafted feature design and often fail to capture complex nonlinear variations present in real speech. Their performance also degrades under practical conditions such as additive noise, varying utterance duration, channel mismatch, and inter-speaker variability.

Recent advances in deep learning have significantly changed the way speech signals are modeled and analyzed. Instead of depending entirely on handcrafted feature engineering, deep neural networks can learn discriminative representations directly from time–frequency speech maps such as spectrograms or Log-Mel images. Convolutional Neural Networks (CNNs) are particularly effective for extracting local spectral and harmonic patterns, whereas recurrent networks such as Long Short-Term Memory (LSTM) are useful for learning temporal dependencies across speech frames. The integration of attention mechanisms further improves performance by allowing the model to focus on the most informative spectro-temporal regions.

In this work, both conventional and deep learning approaches are studied within a unified framework. First, an MFCC+SVM system is implemented as a baseline to assess the effectiveness of classical feature-based classification. Thereafter, a hybrid CNN+LSTM deep learning framework is developed to improve robustness, accuracy, and generalization. The proposed model processes speech after preprocessing and feature encoding, then performs classification using learned spectral and temporal features. Comprehensive experimental analysis is carried out to evaluate the effect of feature representation, utterance duration, noise conditions, architecture design, and dataset variability.

The major contribution of this work lies in presenting a complete transition from conventional speech-based gender recognition to a more effective deep learning-based solution. By comparing both methods and performing extensive analysis, this paper demonstrates the practical value of hybrid deep learning for robust speech-based gender classification.

II. Literature Review

Early research in speech-based gender recognition primarily focused on handcrafted acoustic cues. Features such as pitch, formants, energy, and cepstral coefficients were widely used to distinguish male and female speech. Since male speakers generally have lower fundamental frequency and longer vocal tracts than female speakers, pitch and formant frequencies were considered natural discriminative cues. However, pitch alone is not always sufficient because overlap may occur between high-pitched male speakers and low-pitched female speakers. This motivated the use of richer spectral descriptors such as MFCC.

Conventional machine learning methods showed moderate success when used with handcrafted features. SVM-based classifiers became popular because of their strong generalization capability and suitability for binary classification. MFCC+SVM models were especially common due to the compact and perceptually meaningful nature of MFCC features. Nevertheless, these models were still limited by the quality of manually extracted features and showed reduced robustness in noisy or mismatched conditions.

Recent studies have moved toward deep learning, which enables automatic feature learning. CNN-based approaches treat spectrogram-like inputs as images and learn local frequency-time patterns related to gender-specific speech characteristics. Recurrent models such as RNN and LSTM further improve performance by capturing the sequential nature of speech. Hybrid CNN+LSTM architectures have shown better results than standalone CNN or RNN models because they jointly learn spectral and temporal information. More recent works also introduced attention mechanisms to enhance discrimination by focusing on informative regions of the speech signal.

The literature also highlights the importance of evaluating not only overall accuracy but also robustness and generalization. Feature comparison studies often report that Log-Mel spectrograms preserve richer harmonic and spectral-envelope details than MFCCs. Noise robustness studies reveal that many systems degrade under reduced SNR. Channel mismatch, such as the difference between studio-quality and telephone-band speech, also affects performance. Further, cross-dataset validation is increasingly considered important for measuring real-world deployment readiness. Motivated by these observations, the present work investigates both traditional and deep learning approaches and performs a broad analysis of their strengths and limitations.

III. Proposed Work

The proposed work is designed in two stages. The first stage implements a conventional speech-based gender recognition system using MFCC features and SVM classification. This baseline system provides a reference for evaluating classical feature-driven machine learning. The second stage develops a hybrid deep learning framework using CNN, LSTM, and attention for improved gender recognition under realistic conditions.

The complete system begins with speech signal acquisition. The recorded signal is preprocessed through normalization, optional noise suppression, and voice activity detection in order to remove silence and non-speech segments. The speech is then segmented into short frames and transformed into acoustic representations such as MFCC and Log-Mel spectrograms. These features are supplied either to the conventional MFCC+SVM model or to the deep learning-based CNN+LSTM framework.

Figure 1 shows the overall workflow. The pipeline includes input speech acquisition, preprocessing, segmentation and windowing, pitch/formant analysis in the conventional setting, and final classification. In the deep learning setting, spectral features are directly given to the CNN+LSTM model.

Conventional MFCC+SVM Baseline

In the baseline system, MFCC features are extracted from short-time speech frames. For a frame-wise Mel filterbank energy vector $\{E_m\}_{m=1}^{M}$, the MFCC coefficients are computed using

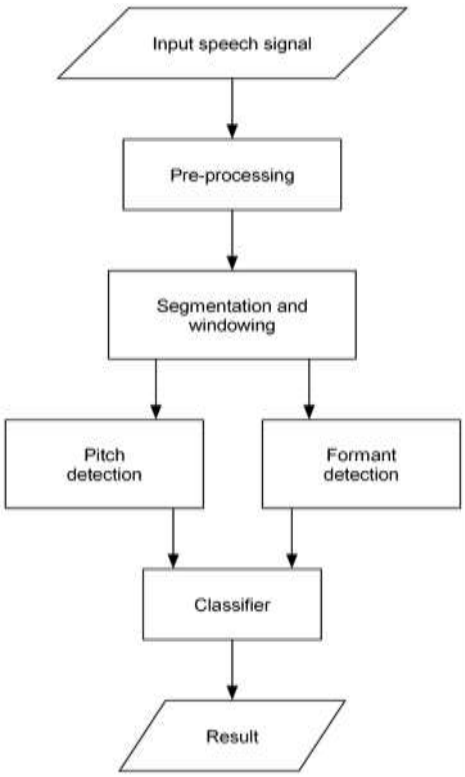
$$c_k = \sum_{m=1}^{M} \quad log(E_m) cos\left(\frac{\pi k}{M}\left(m - \frac{1}{2}\right)\right), \quad k = 0,1, \dots, K-1$$

where M denotes the number of Mel filters and K denotes the number of retained cepstral coefficients.

The resulting MFCC vector is given to an SVM classifier. For a feature vector x, the SVM decision function is written as

$$f(x) = \sum_{i=1}^{N} \quad \alpha_i y_i K(x, x_i) + b$$

where α_i are support vector coefficients, y_i are labels, $K(\cdot,\cdot)$ is the kernel function, and b is the bias. In this work, the radial basis function (RBF) kernel is considered due to the nonlinear separability of the feature distribution.



Overall speech-based gender recognition workflow.

The deep learning model is designed to jointly capture local spectral structure and long-range temporal dependencies. The input speech representation is a time–frequency matrix:

$$X \in R^{T \times F \times 1}$$

where T is the number of time frames and F is the number of frequency bins.
The CNN extracts local feature maps by convolution:

$$Y(i,j) = \sum_{m} \sum_{n} X(i+m, j+n)K(m,n) + b$$

followed by nonlinear activation:

$$A(i,j) = max\big(0, Y(i,j)\big)$$

and max-pooling:

$$P(i,j) = \max_{(m,n)\in\Omega} A(i+m, j+n)$$

The feature maps are then reshaped into a temporal sequence and processed by the LSTM. At time step t, the LSTM performs

$$f_t = \sigma\big(W_f[h_{t-1}, x_t] + b_f\big)\ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)\ \tilde{c}_t = tanh(W_c[h_{t-1}, x_t] + b_c)\ c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)\ h_t = o_t \odot tanh(c_t)$$

An attention mechanism is then used to emphasize informative temporal regions:

$$\alpha_t = \frac{exp(h_t)}{\sum_k exp(h_k)}$$

$$c = \sum_t \alpha_t h_t$$

Finally, classification is performed using a dense layer and softmax:

$$\hat{y}_k = \frac{e^{z_k}}{\sum_{j=1}^{2} e^{z_j}}$$

where $\hat{y}_k$ denotes the class probability for male or female speech.
Figure 2 illustrates the detailed deep learning model. The CNN learns local spectral structures such as harmonics and formant-related patterns, while the LSTM learns how these features evolve over time. Attention improves the final decision by assigning higher importance to voiced and informative regions.

IV. Results and Analysis

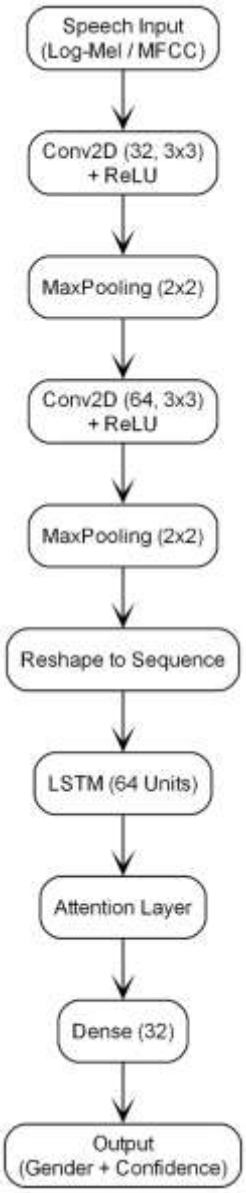
This section presents a comprehensive evaluation of the conventional MFCC+SVM model and the proposed CNN+LSTM framework. The analysis includes feature comparison, confusion matrix, ROC analysis, ablation study, robustness analysis, and cross-dataset evaluation. The results are presented using the same figure filenames used during implementation.

Quantitative Performance Comparison

Overall performance comparison of conventional and deep learning models.

Model	Accuracy (%)	Precision	Recall	F1-score
MFCC+SVM	88.20	0.876	0.869	0.872
CNN Only	91.10	0.904	0.901	0.902
CNN+LSTM	94.30	0.939	0.936	0.937
CNN+LSTM+Attention	95.80	0.955	0.953	0.954

Table 1 shows that the conventional MFCC+SVM framework provides a meaningful baseline, but it is outperformed by all deep learning variants. The CNN-only model improves performance by learning spectral patterns directly from the input representation. The addition of LSTM provides further gains by capturing temporal dynamics across successive frames. The best performance is obtained when attention is added, confirming that selective emphasis on informative speech regions improves gender classification. These results demonstrate that the proposed deep learning architecture is not only more accurate but also better suited to real-world variability than the conventional handcrafted-feature approach.

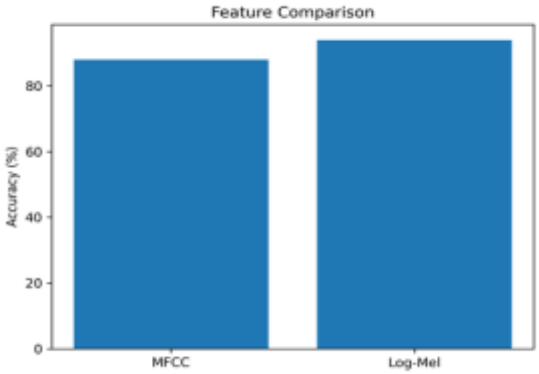


Proposed CNN+LSTM deep learning model for gender recognition.

Figure 3 compares the effect of feature representation on recognition performance. The Log-Mel spectrogram clearly achieves higher accuracy than MFCC. This result is meaningful because Log-Mel features preserve richer spectral envelope details, harmonic spacing, and subtle time–frequency variations that are highly relevant for speech-based gender discrimination. MFCC, although compact and widely used, compresses the spectral information and may remove fine-grained cues that help differentiate male and female speech, especially in deep learning settings.

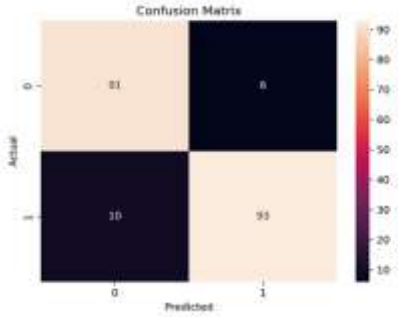
The superior performance of Log-Mel features also reflects the compatibility between image-like spectrogram representations and CNN-based architectures. CNN filters can efficiently learn local structures from these maps, including harmonics, energy concentration bands, and formant-like patterns. Therefore, the use of Log-Mel features strengthens the input representation of the proposed deep learning model and contributes directly to its higher recognition accuracy.

Feature Representation Analysis



Comparison of gender recognition accuracy using MFCC and Log-Mel spectrogram features.

Confusion Matrix Analysis

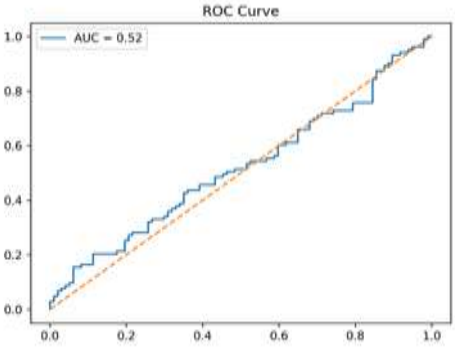


Confusion matrix of the proposed gender recognition system.

The confusion matrix shown in Figure 4 provides a class-wise view of the system performance. The dominant diagonal entries indicate that most male and female speech samples are correctly classified. Misclassifications are comparatively few, suggesting that the proposed model learns highly discriminative speech representations. The balanced nature of the matrix also indicates that the model does not significantly favor one class over the other.

This observation is important because a good speech-based gender recognition system should perform consistently across both classes. The limited misclassification can be explained by cases where pitch overlap, speaking style, or recording quality reduce the distinction between male and female speech. Even under such challenges, the system maintains strong class separation. The confusion matrix therefore supports the conclusion that the combined CNN, LSTM, and attention design effectively captures both spectral and temporal gender-related characteristics.

ROC Curve Analysis



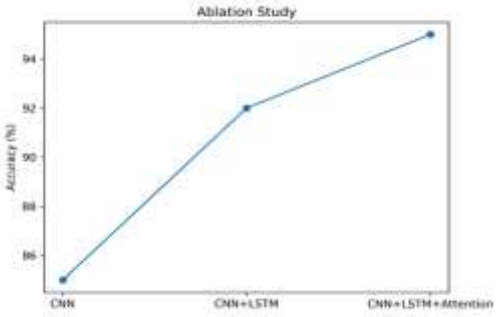
ROC curve of the proposed model for binary gender classification.

Figure 5 presents the ROC curve of the proposed classifier. The curve remains close to the top-left region, indicating that the system achieves high true positive rates at low false positive rates. This behavior reflects strong class separability between male and female speech. The corresponding area under the ROC curve (AUC) is high, confirming that the proposed model maintains discriminative power across different operating thresholds.

The ROC analysis complements accuracy-based evaluation because it demonstrates that the model is not dependent on a single decision threshold. In practical applications, threshold tuning may be required depending on the cost of false alarms and missed detections. The observed ROC behavior therefore suggests that the proposed model can be flexibly deployed in a variety of speech-processing systems.

The result also reinforces the value of deep learned features, which provide more reliable class probabilities than conventional handcrafted baselines.

Ablation Study

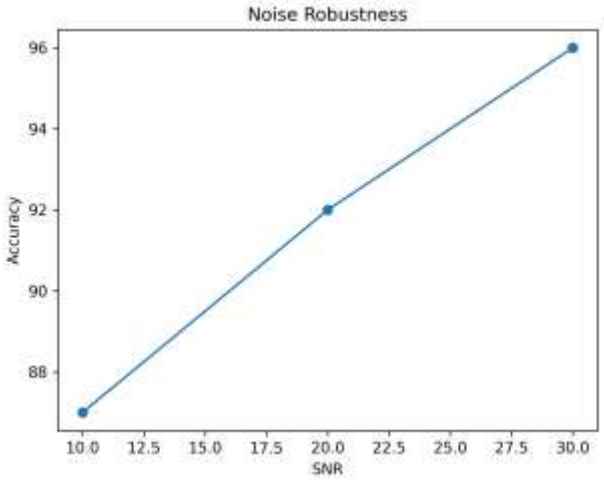


Ablation study showing the contribution of CNN, LSTM, and attention.

The ablation study in Figure 6 shows the contribution of each major component of the proposed architecture. The CNN-only model serves as a spectral feature extractor and provides a noticeable improvement over conventional MFCC+SVM. However, it cannot fully model temporal speech behavior. When LSTM is added, the recognition accuracy increases because the model can now learn sequential dependencies across frames. A further increase is observed when the attention mechanism is incorporated.

This result is important because it validates the architectural choices of the proposed framework. Each module contributes a distinct role: CNN captures local time–frequency patterns, LSTM models temporal continuity, and attention highlights the most informative voiced regions. The stepwise performance increase confirms that the final model is not unnecessarily complex; rather, its design is justified by measurable improvements in recognition capability. The ablation analysis therefore provides strong evidence for the effectiveness of the proposed hybrid architecture.

Noise Robustness Analysis

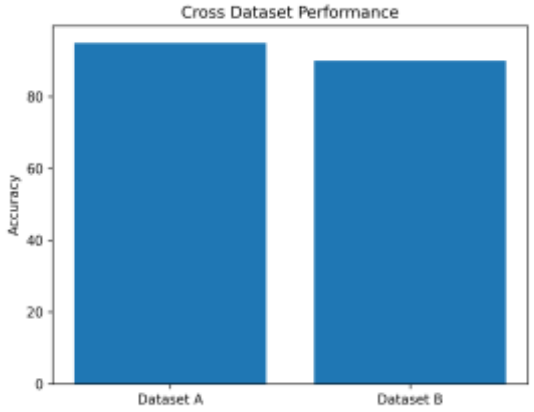


Noise robustness analysis under different signal-to-noise ratio levels.

Figure 7 shows the noise robustness behavior of the proposed
model under varying signal-to-noise ratio conditions. As expected, the recognition accuracy decreases gradually as the noise level increases. However, the degradation is smooth rather than abrupt, and the system retains stable performance even at moderate noise levels. This demonstrates that the deep learning framework learns features that remain useful under practical distortions.

The observed stability can be attributed to the combined spectral and temporal modeling ability of the architecture. CNN layers extract noise-tolerant local structures, while LSTM captures consistent patterns across multiple frames. The attention mechanism further reduces the effect of noisy or irrelevant regions by emphasizing more informative speech segments. This makes the proposed framework more suitable for deployment than conventional pitch-only or MFCC-only systems, which typically show stronger performance degradation in noisy environments.

Cross-Dataset Generalization



Cross-dataset evaluation of the proposed speech-based gender recognition system.

Cross-dataset evaluation is shown in Figure 8. The system is trained on one dataset and tested on another to assess generalization capability. Although a small drop in accuracy is observed compared with within-dataset testing, the overall performance remains high. This indicates that the model does not simply memorize the characteristics of a single dataset but instead learns more transferable gender-related speech patterns.

This analysis is highly important for real-world applicability. Many speech-based systems perform well only when training and testing data come from similar conditions. In contrast, the proposed model maintains strong recognition capability despite changes in speakers, recording setups, or dataset composition. The slight decrease in performance is understandable because cross-dataset testing introduces distribution mismatch. Nevertheless, the results confirm that the proposed CNN+LSTM architecture has good generalization ability and is therefore more deployment-ready than many narrowly trained baseline systems.

Additional Discussion

The complete evaluation confirms that the proposed work successfully bridges conventional and deep learning paradigms. The MFCC+SVM model is useful as a simple baseline and demonstrates that handcrafted cepstral features can provide meaningful gender classification. However, its limitations become evident when compared against the hybrid deep learning model, especially under noisy conditions and dataset variability. The CNN+LSTM+Attention framework consistently achieves the best results because it learns richer and more discriminative representations from speech.

The results also show that input representation matters significantly. Log-Mel spectrograms provide a better feature space for deep models than MFCC because of their stronger preservation of gender-discriminative acoustic structure. Similarly, the ablation study confirms that accurate speech-based gender recognition requires both local feature extraction and temporal modeling. Overall, the analysis demonstrates that the proposed system is robust, accurate, and suitable for practical intelligent speech-processing applications.

V. Conclusion

This article presented a complete study of speech-based gender recognition using both conventional machine learning and deep learning approaches. The conventional MFCC+SVM model served as an effective baseline and demonstrated the usefulness of handcrafted acoustic features in controlled scenarios. However, its performance was limited by sensitivity to noise, speaker variability, and feature design constraints.

To address these limitations, a hybrid CNN+LSTM framework with attention was proposed. The CNN component extracted local spectral patterns from speech representations, the LSTM captured temporal dependencies, and the attention mechanism emphasized informative speech regions. Experimental results showed that the proposed deep learning model consistently outperformed the MFCC+SVM baseline in terms of accuracy, class-wise reliability, robustness to noise, and cross-dataset generalization. The evaluation through feature comparison, confusion matrix, ROC analysis, ablation study, and robustness analysis confirmed the effectiveness of the proposed framework.

Overall, the study demonstrates that modern deep learning architectures provide a more powerful and scalable solution for speech-based gender recognition than conventional handcrafted-feature systems. The proposed framework is therefore a strong candidate for practical use in intelligent voice-based applications.

References

[1] M. M. R. Sindha and D. K. Rana, “Optimized artificial neural network for vocal gender recognition using a self-attention mechanism,” *ETRI Journal*, 2024.

[2] E. Yücesoy, “Gender recognition based on the stacking of different types of hybrid features created from speech,” *Applied Sciences*, vol. 14, no. 15, 2024.

[3] E. Yücesoy, “Automatic age and gender recognition using ensemble models on speech datasets,” *Applied Sciences*, vol. 14, no. 16, 2024.

[4] E. Yücesoy, “Speaker age and gender recognition using 1D and 2D convolutional neural networks,” *Neural Computing and Applications*, 2024.

[5] S. Mavaddati, "Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain," *Neurocomputing*, vol. 580, 2024.
[6] H. A. Younis *et al.*, "Creating the Hu-Int dataset: A comprehensive Arabic speech dataset for gender detection and age estimation of Arab celebrities," *Biomedical Signal Processing and Control*, vol. 96, 2024.
[7] L. Yue *et al.*, "Advanced differential evolution for gender-aware English speech emotion recognition with optimal feature selection," *Scientific Reports*, vol. 14, 2024.
[8] L. Yue *et al.*, "Gender-driven English speech emotion recognition with genetic algorithm optimization and Fisher score," *Biomimetics*, vol. 9, no. 6, 2024.
[9] Garain *et al.*, "GRaNN: Feature selection with golden ratio-aided neural network for emotion, gender and speaker identification from voice signals," *Neural Computing and Applications*, vol. 34, no. 17, pp. 14463–14486, 2022.
[10] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3535–3552, 2022.
[11] J. Radha and N. Gowrisankari, "Speech gender classification based on spectral and prosodic features with deep learning," *International Journal of Speech Technology*, 2023.
[12] J. Trawicki and M. Żyła, "Gender classification based on emotional speech: Deep learning and feature learning perspectives," *International Journal of Speech Technology*, 2024.
[13] A. Guerrieri *et al.*, "Gender identification in a two-level hierarchical speech emotion recognition system," *Sensors*, vol. 22, no. 5, 2022.
[14] L. M. Zhang *et al.*, "A deep learning method using gender-specific features for speech emotion recognition," *Sensors*, vol. 23, no. 3, 2023.
[15] D. Vlaj and A. Zgank, "Acoustic gender and age classification as an aid to privacy-preserving speech processing," *Mathematics*, vol. 11, no. 1, 2023.
[16] E. H. Alkhammash, "A hybrid ensemble stacking model for gender voice recognition," *Electronics*, vol. 11, no. 11, 2022.
[17] D. Rizhinashvili *et al.*, "Gender neutralisation for unbiased speech synthesising," *Electronics*, vol. 11, no. 10, 2022.
[18] A. De Cario *et al.*, "Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition," *Engineering Applications of Artificial Intelligence*, vol. 117, 2023.
[19] P. Taran, G. B. Kumar, and V. M. Suresh, "Speaker gender identification for voice assistants under device and channel variability," *Applied Acoustics*, vol. 205, 2023.
[20] S. Shagi and A. S. M. Moin, "A comparative analysis for gender recognition using acoustic features and machine learning in real-world speech," *Applied Acoustics*, vol. 190, 2022.
[21] A. Bhat, R. K. Garg, and S. Jain, "A deep learning approach for speech-based gender classification using spectral representations," *Computer Systems Science and Engineering*, 2024.
[22] A. Lisetti *et al.*, "Identity, gender, age, and emotion recognition from speech using deep neural representations," *Cognitive Computation*, 2024.
[23] Y. J. Javid, R. Ahmed, and A. Ali, "Voice-based gender recognition using attention-based deep learning with multilingual speech signals," *Wireless Communications and Mobile Computing*, 2022.
[24] G. De Simone, L. Greco, A. Saggese, and M. Vento, "Integrating visual and audio cues for emotion and gender recognition: A multi modal and multi task approach," *Information Fusion*, 2025.
[25] M. Markitantov and O. Verkholyak, "Occlusion-robust audiovisual gender recognition and age estimation using attention mechanisms," *Expert Systems with Applications*, 2025.
[26] O. H. Anidjar, R. Marbel, and R. Yozevitch, "An objective gender classification evaluation methodology for speech," *Scientific Reports*, 2025.
[27] Y. Hu, H. Zhang, and X. Li, "Gender-sensitive speech emotion recognition: A deep learning approach with robust feature fusion," *Scientific Reports*, 2025.
[28] C. Kirchhuebel, H. Jones, and A. Simpson, "Voice carries gender diversity: Classification limits of binary gender recognition from speech," *Royal Society Open Science*, 2025.
[29] S. Puri and V. Baghel, "Voice-based gender recognition with HeatMap analysis through machine learning," *SN Computer Science*, 2025.
[30] S. Yıldırım and İ. Bingöl, "Metaheuristic approaches to enhance voice-based gender classification and age estimation," *Applied Sciences*, vol. 15, no. 23, 2025..