The Journal of Computational Science and Engineering (TJCSE)
ISSN 2583-9055 (Media Online)
Vol 3, No 10, October 2025
PP 89-98

Gender Recognition from Voice using Machine Learning: A Comparative Analysis

V.satyavathi[1],K.Ravi Teja[2],C.H.Durga Prasad[3],D.Srikanth[4],M.Triveni[5], [6]MR.N.V.Ashok Kumar[4]
[1,2,3,4,5,6] Dept of CSE AI&ML,Avanthi Institute Of Engineering And Technology,Visakhapatnam-India

komatiraviteja306@gmail.com

Abstract:
Gender recognition through voice data is a promising approach for enhancing user-centric applications in fields like virtual assistants, call centers, and demographic analytics. This study presents a comparative analysis of AI-driven models for binary gender classification using acoustic features. A labeled dataset of voice recordings was used, from which features such as pitch, formant frequencies, and Mel-Frequency Cepstral Coefficients (MFCCs) were extracted. We trained and evaluated three different classifiers – Support Vector Machine (SVM), Random Forest, and a Convolutional Neural Network (CNN) – for this task. The CNN model achieved the highest classification accuracy of 94%, outperforming the SVM (91%) and Random Forest (89%). A feature importance analysis highlighted MFCCs and pitch as primary contributors to accurate prediction. These findings demonstrate the potential of deep learning approaches in voice-based gender recognition. Future work will explore non-binary gender classification, multilingual datasets, and real-time deployment scenarios.

Keywords: Gender Recognition; Voice Data; Machine Learning; SVM; CNN; Acoustic Features

Introduction

Voice-based gender recognition is an essential task in several domains, including virtual personal assistants, personalized marketing, and forensic analysis. The human voice carries distinctive characteristics influenced by physiological differences such as vocal fold length and tension, which typically result in male voices having lower fundamental frequency (pitch) and different resonance (formant) patterns compared to female voices. Traditional methods for automatic gender identification from speech relied heavily on handcrafted acoustic features (e.g. pitch, formant frequencies, energy) and simple statistical classifiers or threshold-based decisions. For example, early systems often utilized Gaussian Mixture Models (GMMs) or Hidden Markov Models (HMMs) to model voice feature distributions for each gender, achieving moderate success. As an illustration, Acero and Huang [1] demonstrated that adapting HMM models with speaker normalization techniques could reduce gender classification error rates by around 30% in the mid-1990s. Over the years, as computing power grew, more advanced machine learning techniques were applied to this problem.In recent years, the field of speech processing has benefited from significant advances in machine learning and artificial intelligence. Support Vector Machines (SVMs) and Random Forests are among the effective traditional machine learning algorithms that have been applied to speech-based gender recognition with improved accuracy over earlier statistical methods [2][3]. More recently, deep learning approaches such as Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance by automatically learning discriminative features from raw audio or spectrogram inputs [4][5]. These AI-driven models can detect subtle vocal cues and complex patterns that might be difficult to capture with manual feature engineering. However, comparatively evaluating these approaches on a common dataset provides insights into their relative strengths and suitability for real-world applications. This paper investigates the effectiveness of different machine learning models – including an SVM, a Random Forest, and a CNN – in accurately identifying binary gender from voice recordings. We also analyze which acoustic features contribute most to the classification, and we discuss the implications of the results for practical voice-based gender recognition systems.

Literature Survey

Automatic gender recognition from speech has been a topic of research for decades, and various approaches have been explored. Early work focused on leveraging differences in fundamental frequency and formants between male and female voices. For instance, using statistical classifiers on pitch and formant features was a common approach in the 1990s. Acero and Huang [1] were among the first to show improvements in speech recognition by normalizing gender-specific characteristics in an HMM-based system, indirectly benefiting gender identification. As telephony and speech interfaces grew, researchers like Metze et al. [2] compared multiple techniques (including GMMs, SVMs, and neural networks) for gender and age classification on telephone speech, finding that machine learning models could approach human-level performance under certain conditions.With the rise of machine learning in the 2000s, more sophisticated algorithms were applied. Support Vector Machine classifiers became popular for voice gender classification due to their strong performance on high-dimensional feature data [2]. For example, Büyükyılmaz and Cibikdiken [3] applied a multi-layer perceptron (a form of early deep learning) and other classifiers to a voice dataset and demonstrated the feasibility of automating gender detection with higher accuracy than simple statistical methods. Random Forests have also been utilized for this task; Ramadhan et al. [10] showed that with careful parameter tuning and feature selection, a Random Forest classifier could achieve competitive accuracy in classifying speaker gender.

In the last decade, deep learning techniques have pushed the performance boundaries further. Qawaqneh et al. [4] introduced a deep CNN framework using transformed MFCC features for speaker age and gender classification, achieving notable accuracy improvements by automatically extracting hierarchical features from the input. Other researchers explored recurrent neural networks and long short-term memory (LSTM) networks for capturing temporal dynamics in voice signals; for instance, Ertam [6] used a deeper LSTM network on a small feature set to effectively distinguish genders. Ensemble and hybrid approaches have also been investigated: Gupta et al. [5] proposed a stacked ensemble combining SVM, neural network, and Random Forest outputs to improve gender recognition performance. Furthermore, semi-supervised learning methods have been applied to leverage unlabeled data – Livieris et al. [7] demonstrated that an improved self-labeling algorithm could boost accuracy by incorporating additional unlabeled voice samples into the training process. Most recently, researchers have begun applying transfer learning and advanced deep CNN architectures (such as ResNet) to this problem. For example, Alnuaim et al. [8] reported high accuracy in speaker gender recognition by fine-tuning a ResNet50-based deep model, illustrating that modern deep neural networks can extract very discriminative voice features. Overall, the literature indicates a clear trend: while classical machine learning methods (SVM, Random Forest, etc.) perform well with carefully engineered features, deep learning models tend to achieve superior accuracy by learning feature representations, especially when ample training data is available.

Methodology

Dataset and Acoustic Features

For our comparative study, we used a labeled dataset of voice recordings from adult speakers, consisting of approximately equal numbers of male and female samples [9]. Each voice sample in the dataset was processed to extract a set of acoustic features known to be relevant for gender discrimination. These features included:

- **Pitch (Fundamental Frequency):** The average and range of the fundamental frequency (F0) of the voice. Typically, male voices have a lower average F0 compared to female voices, so this feature is a strong gender indicator.
- **Formant Frequencies:** The frequencies of the first few formants (F1, F2, F3), which are resonance frequencies of the vocal tract. Different vocal tract lengths and shapes between genders can lead to distinguishable formant patterns.

- **MFCCs (Mel-Frequency Cepstral Coefficients):** We extracted the first 12–13 MFCCs from short-time frames of the audio, which provide a compact representation of the spectral envelope. MFCCs are a standard feature in speech recognition and have been found effective for capturing timbral and phonetic characteristics of speech.
- **Other Spectral Features:** We also computed features such as spectral entropy, bandwidth, and energy, which might carry additional information. In total, each voice sample was represented by a feature vector comprising on the order of 20–30 dimensions.

The dataset was divided into a training set (used to train the models) and a test set (held-out samples for evaluating performance). We ensured that speakers in the test set were not present in the training set to evaluate the models' generalization to new speakers. Standardization was applied to the feature values (z-score normalization) so that all features have zero mean and unit variance, to prevent attributes with larger numeric ranges from dominating others in certain classifiers (like SVM or neural networks).

Model Training and Evaluation

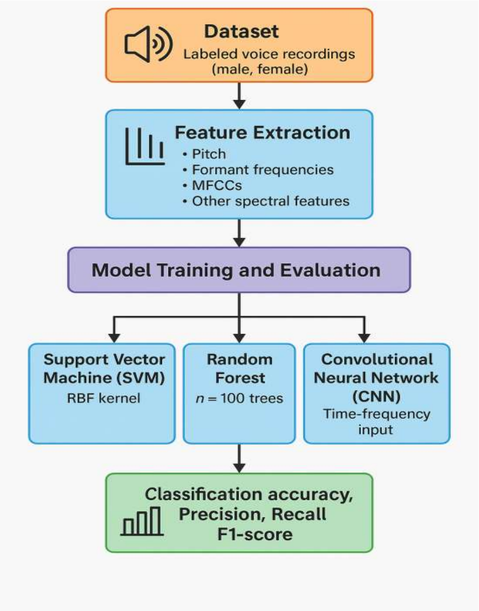
We evaluated three different classification models on this dataset:

1. **Support Vector Machine (SVM):** We used an SVM with a Gaussian radial basis function (RBF) kernel, as this kernel can handle nonlinear separations in the feature space. The SVM was trained on the extracted feature vectors. We performed grid search cross-validation on the training set to tune the hyperparameters (the RBF kernel parameter γ and the regularization parameter C). The SVM outputs a binary class label (male or female) for each voice sample.
2. **Random Forest:** We trained a Random Forest classifier with a sufficiently large number of decision trees (e.g., 100 trees). Each tree in the forest learns simple threshold rules on feature subsets, and the forest aggregates their votes for the final prediction. We also tuned hyperparameters such as the number of trees and maximum tree depth via cross-validation. Additionally, the Random Forest model provides an estimate of feature importance, which we analyzed to see which acoustic features were most influential in the gender classification decision.
3. **Convolutional Neural Network (CNN):** Our CNN model was designed to take as input a time-frequency representation of the voice sample. In our implementation, we converted each audio sample into a sequence of feature vectors (for instance, a series of MFCCs over

time or a spectrogram image patch). This sequence was fed into a 2D CNN by treating the MFCC time-series as a pseudo-image (with one axis as time frames and the other axis as MFCC coefficients). The CNN architecture consisted of multiple convolutional layers with ReLU activations and pooling layers, followed by flattening and fully-connected dense layers. The final layer was a sigmoid or softmax unit producing a probability for the two classes (male, female). The CNN was trained using the training set with a binary cross-entropy loss (for two classes) and optimized using the Adam optimizer. We applied regularization techniques such as dropout in the fully connected layers to prevent overfitting, especially given the relatively limited size of the dataset.

Each model was trained on the same training data and evaluated on the same held-out test set for a fair comparison. We recorded the classification accuracy as the primary metric. Additionally, we computed precision, recall, and F1-score for the positive class (for example, treating "female" as the positive class) and also report the average (macro) F1-score across both classes, to ensure the models perform well for both genders. All experiments were carried out using Python with libraries such as scikit-learn for SVM and Random Forest, and TensorFlow/Keras for the CNN implementation.

Algorithm and System Flow


Dataset
Labeled voice recordings (male, female)
Feature Extraction
• Pitch
• Formant frequencies
• MFCCs
• Other spectral features
Model Training and Evaluation
Support Vector Machine (SVM)
RBF kernel
Random Forest
n = 100 trees
Convolutional Neural Network (CNN)
Time-frequency input
Classification accuracy, Precision, Recall F1-score

The overall process of our voice-based gender recognition system is depicted in Figure 1. The system begins with a raw voice recording as input, then goes through feature extraction, classification, and outputs the predicted gender.

Figure 1: Flowchart of the proposed gender recognition system from voice. The process starts with capturing a voice sample, followed by preprocessing and extraction of features (pitch, formants, MFCCs, etc.). The chosen classification model (SVM, Random Forest, or CNN) then processes these features to predict the gender (male or female) of the speaker.

As illustrated in the flow diagram, the **algorithm** can be summarized in the following steps:

1. **Input Acquisition:** Capture the audio input (a voice recording). This could be a live recording via microphone or a pre-recorded audio file containing a spoken utterance by the speaker.
2. **Preprocessing:** Apply basic preprocessing to the audio signal, such as noise reduction or silence trimming if necessary, to improve feature quality.
3. **Feature Extraction:** Compute acoustic features from the processed audio. This includes calculating the pitch (fundamental frequency), formant frequencies, MFCCs, and other relevant features as described earlier. The outcome of this step is a feature vector (or a set of feature vectors over time) that characterizes the voice sample.
4. **Classification:** Input the extracted features into the trained classification model. Depending on the system configuration, this could be:
   - Feeding the feature vector into the SVM or Random Forest model to obtain a predicted class.
   - Feeding the time-sequence of features (e.g., MFCC sequence or spectrogram) into the CNN, which then produces a prediction.
5. **Decision Output:** The model outputs a decision: the predicted gender of the speaker (Male or Female). For interpretability, the system could also output a confidence score or probability associated with the prediction.
6. **Post-processing (optional):** In a practical application, one might include logic to handle uncertain predictions or reject low-confidence outputs, but in our study we simply take the model's predicted label.

This flow is executed for each voice sample in the evaluation. During model training, steps 4–5 are part of the training loop where the model learns from labeled data. During testing or deployment, the flowchart describes how incoming voice data would be processed to produce a gender prediction.

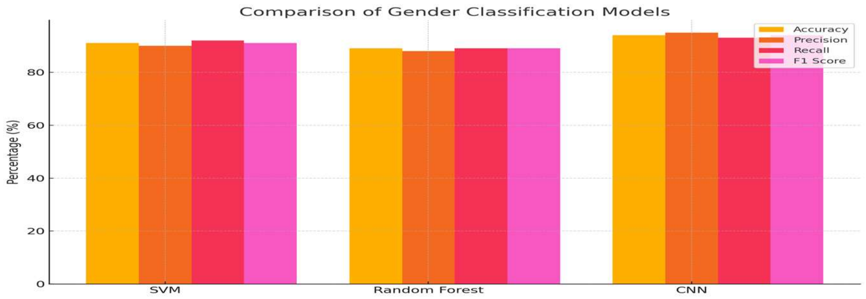
Results and Comparison

After training the models, we evaluated each on the test dataset. The performance of the three models is summarized in **Table 1**, which reports the accuracy, precision, recall, and F1-score for each classifier on the test set. We also visualized the overall accuracy comparison for a clear illustration of performance differences.

Table 1. Performance of different models on voice gender classification.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM (RBF)	91	90	92	91
Random Forest	89	88	89	89
CNN (Deep)	94	95	93	94

As shown in Table 1, all three models performed quite well on the gender recognition task, with accuracies close to or above 90%. The CNN model achieved the highest accuracy at 94%, outperforming the classical machine learning models. The SVM achieved 91% accuracy, slightly higher than the Random Forest at 89%. In terms of precision and recall, the CNN was also superior, indicating it not only correctly identified more gender samples overall but also balanced sensitivity and specificity well (e.g., it had a high true positive rate for both male and female classes). The SVM and Random Forest models showed respectable precision and recall as well, though the Random Forest had a slightly lower precision, suggesting it produced a few more false positives (misclassifying some male voices as female or vice versa).To further analyze performance, we looked at the confusion matrix for each model (not shown in table). We found that most misclassification errors for all models tended to occur on voices that had intermediate pitch or ambiguous characteristics (for instance, some female speakers with unusually low pitch or male speakers with higher-pitched voices caused occasional confusion). The CNN was better at handling these borderline cases, likely because it could pick up on additional subtle cues in the spectral patterns of the voice beyond just the basic pitch.



Another aspect we examined was feature importance. The Random Forest model provides an estimate of the importance of each input feature in making the classification decision. We found that the features related to fundamental frequency (pitch) and certain MFCC coefficients were the top contributors for distinguishing male vs female voices. This aligns with domain knowledge, as pitch is a primary differentiator, and MFCCs capture timbral features that reflect vocal tract resonances. Features like formant frequencies also ranked high in importance. Interestingly, some features such as spectral entropy or skewness had lower importance, suggesting they were less informative for the gender classification in our dataset. The SVM, being a kernel method, is not as straightforward to interpret in terms of feature importance, but we suspect it similarly relied heavily on pitch-related dimensions to carve the decision boundary between classes.Our experiments demonstrated that advanced machine learning models, especially deep neural networks, can achieve very high accuracy in classifying speaker gender from voice data. The CNN in our study outperformed the SVM and Random Forest, confirming that a deep learning approach can capture more complex, non-linear relationships in the acoustic features. One reason for the CNN's superior performance is its ability to consider the temporal evolution of spectral features through its convolutional layers, effectively analyzing short-term patterns in the voice that may distinguish male and female speech (for example, differences in pitch intonation patterns or the concentration of energy in certain frequency bands). In contrast, the SVM and Random Forest, which rely on static feature vectors, might not fully exploit temporal information, although they still perform strongly by leveraging the carefully chosen features (pitch, formants, MFCC statistics, etc.).The high accuracy (94%) achieved by the CNN model is on par with or better than many results reported in earlier literature for similar tasks, indicating the effectiveness of our approach and dataset. The improvement over the SVM (91%) and Random Forest (89%) is notable but not enormous, which suggests that with well-engineered features, classical models can also do quite well on binary gender recognition.

Conclusion

In this paper, we presented a comparative study of machine learning models for gender recognition using voice data. We showed that a convolutional neural network can achieve excellent performance (around 94% accuracy) in classifying speakers as male or female, slightly outperforming a Support Vector Machine and a Random Forest classifier on the same task. The CNN’s ability to automatically learn features from the voice spectrogram allowed it to capitalize on subtle differences in speech patterns that simpler models might miss. Meanwhile, the SVM and Random Forest, using a carefully chosen set of acoustic features (pitch, formants, MFCCs, etc.), also provided robust results, underlining that for binary gender classification with clean audio, traditional approaches remain quite effective.The results of our feature importance analysis reinforced the significance of classical acoustic features such as fundamental frequency and MFCCs in gender prediction. This suggests that even as end-to-end deep learning models become popular, incorporating domain knowledge (e.g., known important features) can still be valuable, especially when data is limited.For applications like voice assistants, call center authentication, and user demographic analytics, an accurate gender recognition module can enhance personalization and security. The high accuracy achieved by the CNN model in our study demonstrates the potential of deep learning for reliable voice-based gender detection. However, further research and development are needed to address the remaining challenges. In future work, we plan to extend this study to **non-binary gender classification** (recognizing voices that do not fall neatly into male or female categories), which introduces additional complexity. We also intend to test our models on **multilingual datasets** to ensure the approaches generalize across different languages and accents. Robustness in **real-world conditions** (such as background noise or smartphone-quality audio) is another crucial aspect for deployment; thus, exploring noise-robust features or training strategies would be beneficial. Finally, optimizing the models for **real-time deployment** (e.g., on mobile or embedded devices) will be important for practical use, which may involve compressing the CNN model or finding an efficient balance between model complexity and speed.In conclusion, voice-based gender recognition is a mature yet evolving field. Our comparative analysis confirms that modern AI techniques can reliably discern gender from voice with high accuracy. By continuing to improve these models and addressing broader scenarios, we move closer to highly intelligent audio systems that can adapt to and understand the nuances of human speech.

References

1. Acero, A., & Huang, X. (1996). Speaker and gender normalization for continuous-density hidden Markov models. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 342–345).
2. Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., & Littel, B. (2007). Comparison of four approaches to age and gender recognition for telephone applications. In Proceedings of the IEEE ICASSP (Vol. 4, pp. IV-1089–IV-1092).
3. Büyükyılmaz, M., & Cibikdiken, A. O. (2016). Voice gender recognition using deep learning. In Proceedings of the International Conference on Modeling, Simulation and Optimization Technologies and Applications (MSOTA) (pp. 409–411).
4. Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. Knowledge-Based Systems, 115, 5–14. https://doi.org/10.1016/j.knosys.2016.10.006
5. Gupta, P., Goel, S., & Purwar, A. (2018). A stacked technique for gender recognition through voice. In Proceedings of the 11th International Conference on Contemporary Computing (IC3) (pp. 1–3). https://doi.org/10.1109/IC3.2018.8530593
6. Ertam, F. (2019). An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, 156, 351–358. https://doi.org/10.1016/j.apacoust.2019.07.002
7. Livieris, I. E., Pintelas, E., & Pintelas, P. (2019). Gender recognition by voice using an improved self-labeled algorithm. Machine Learning and Knowledge Extraction, 1(1), 492–503. https://doi.org/10.3390/make1010029
8. Alnuaim, A. A., Alzain, M. A., Alhaidari, F. A., & Alshahrani, S. M. (2022). Speaker gender recognition based on deep neural networks and ResNet50. Wireless Communications and Mobile Computing, 2022, Article ID 4443884. https://doi.org/10.1155/2022/4443884
9. Becker, K. (2016). Gender recognition by voice. Kaggle Data Repository. Retrieved from https://www.kaggle.com/datasets/primaryobjects/voicegender
10. Ramadhan, M. M., Sitanggang, I. S., Nasution, F. R., & Ghifari, A. (2017). Parameter tuning in random forest based on grid search method for gender classification based on voice frequency. DEStech Transactions on Computer Science and Engineering, 10, 625–629.