**Bulletin of Electrical Engineering and Informatics**
Vol. 14, No. 1, February 2025, pp. 328~337
ISSN: 2302-9285, DOI: 10.11591/eei.v14i1.8146 

# Exploring Bengali speech for gender classification: machine learning and deep learning approaches

**Habiba Dewan Arpita[1], Abdullah Al Ryan[1], Md. Fahad Hossain[2], Md. Sadekur Rahman[1], Md Sajjad[3], Nuzhat Noor Islam Prova[4]**

[1]Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh
[2]Department of Computer Science and Engineering, Bangladesh University, Dhaka, Bangladesh
[3]Master of Information and Communications Technology, Western Sydney University, Sydney, Australia
[4]Seidenberg School of CSIS, Pace University, New York, USA

**Article Info**

***Article history:***

Received Jan 5, 2024
Revised Oct 5, 2024
Accepted Oct 17, 2024

***Keywords:***

Deep learning
Gender classification
Machine learning
Mel-frequency cepstral coefficients
Speech recognition

**ABSTRACT**

Speech enables clear and powerful idea transmission. The human voice, rich in tone and emotion, holds unique beauty and significance in daily life. Vocal pitches vary by gender and are influenced by emotions and languages. While people naturally perceive these nuances, machines often struggle to capture these subtle distinctions. Machines may struggle to detect these nuances, but people effortlessly perceive them. This project aims to use various machine learning (ML) and deep learning (DL) techniques to reliably determine an individual's gender from a corpus of Bengali conversations. Our dataset comprises 3185 Bengali speeches, with 1100 delivered by males, 1035 by women, and 1050 by those who identify as third gender. We employed six distinct feature extraction techniques to examine the audio data: roll-off, spectral centroid, chroma-stft, spectral bandwidth, zero crossing rate, and Mel-frequency cepstral coefficients (MFCC). Extreme gradient boosting (XGBoost), support vector machines (SVM), K-nearest neighbors (KNN), decision trees classifier (DTC), and random forest (RF) were employed as the five ML algorithms to comprehensively analyze the dataset. For a full study, we also included 1D convolutional neural networks (CNN) from the DL area. The 1D CNN performed extraordinarily well, exceeding the accuracy of all other algorithms with a stunning 99.37%.



***Corresponding Author:***

Habiba Dewan Arpita
Department of Computer Science and Engineering, Daffodil International University
Dhaka-1216, Bangladesh
Email: habiba15-14042@diu.edu.bd

## 1. INTRODUCTION

Linguistic and nonlinguistic information can be expressed through human speech, including age, gender, emotional state, language accent, and tone. It also provides paralinguistic signals, which give information about the speaker's intent and attention. The diverse frequencies of each individual's voice are shaped by the complexities of their vocal cord model. The complex quality of human speech is an essential component of productive communication [1].

Speech-based gender categorization is the process of automatically classifying individuals into gender categories, such as male, female, or other gender identities, based on qualities observed in their speech. Advancements in speech recognition technology have completely transformed the process of gender recognition, which used to rely on manual analysis. Speech recognition is a game-changing technology that has altered the way people interact with computers and other gadgets by enabling spoken word

communication and diminishing reliance on conventional input techniques. Speech recognition systems are now significantly more precise and efficient thanks to the integration of machine learning (ML) and natural language processing (NLP) techniques [2]. In the domains of artificial intelligence and human-machine interaction, gender detection from speech is significant because it helps personalize voice-activated systems to the demands of the individual user. Applications include tailored marketing that alters information according to gender preferences [3].

In biology, there are three distinct genders: male, female, and a third gender. Individuals who do not identify as either male or female are classified as belonging to the third gender. When it comes to biology, the term "third gender" usually describes people who do not strictly fall into the binary categories of male or female. This is because of abnormalities in the production of crucial chromosomes. We must acknowledge the presence of individuals who identify as a third gender in our culture. By acknowledging the existence of third-gender individuals, we can safeguard their entitlement to autonomy, equality, and respect [4], [5].

According to our finding speech gender classification traditionally focuses on distinguishing between males and females, neglecting the inclusion of the third gender. Therefore, we have seized this chance to incorporate the third gender into Bengali speech gender classification alongside males and females. Third-gender people have quite different and unique speech tones than males and females and we can differentiate them by hearing their voices also. Speech recognition is still a demanding task, especially when working with specific languages like Bengali. Furthermore, before trying any form of speech recognition, it is important to correctly identify the speaker's gender. This work examines "gender classification from Bengali speech" offering an in-depth investigation of deep learning (DL) and ML methods.

The main purpose of this work is to analyze the effectiveness of particular ML and DL algorithms at the confluence of gender and voice recognition. Using a complete evaluation of feature extraction approaches and algorithmic frameworks, our purpose is to illuminate the barriers and opportunities connected to gender-specific voice recognition, specifically regarding Bengali language patterns. To address the potential gender-related influences on speech recognition system accuracy, we employed a range of algorithms, including extreme gradient boosting (XGBoost), random forest (RF), support vector machines (SVM), K-nearest neighbors (KNN), decision trees (DT), and proposed convolutional neural networks (CNN) [6]. But before these techniques could be employed, it was required to extract features from speech so that computers could correctly analyze and comprehend human speech. To turn essential acoustic characteristics such as roll-off, zero crossing rate, Mel-frequency cepstral coefficients (MFCC), spectral centroid, chroma-short-time Fourier transform (STFT), and spectral bandwidth into a numerical form that can be examined, this process must be accomplished. To support applications like voice recognition, speaker recognition, emotion detection, and language modeling, feature extraction is needed for precise and successful speech processing. We mindfully created a well-balanced dataset of 3,185 samples that yielded the following three speaker groups: men, women, and third gender. The inclusion of the third gender category expanded the notion of gender recognition, filling a research gap.

## 2. LITERATURE REVIEW

In this segment, we conducted a literature review on various existing studies in speech recognition and speech gender recognition. Some researchers use ML methodologies employed in speech recognition research. For speaker verification, Safavi *et al.* [7] I-vector-based age group identification accuracy was 82.62%, and its Gaussian mixture model (GMM)-SVM gender identification accuracy was 79.18%. 1100 kids in three age groups (AG1, AG2, and AG3), ranging from kindergarten to grade 10, participated in the study. A dataset with 20 languages was analyzed by Alkhawaldeh [8], the research was able to attain impressive accuracy wiht DL norm-99.97% and SMO-99.7% in gender recognition. Sharma and Mala [9] employed a hybrid SVM and principal component analysis (PCA) technique to attain a 91% accuracy in gender classification and finally enhanced accuracy to 98.42% by incorporating PCA with SVM Hamdi *et al.* [10] detected gender using the Kaggle Arabic natural audio dataset (ANAD). In Arabic speech, gender could be recovered with 96.02% and 95.30% accuracy, respectively, using models such as linear SVM and logistic regression (LR). Spectral subband centroids (SSCs), MFCC, and energy-related features were employed by Guerrieri *et al.* [11] using the EMOVO dataset and achieved an excellent 97.8% gender detection accuracy by adopting a hierarchical approach with GMM, Gaussian regression (GR), and SVM. Gupta *et al.* [12] conducted a comparison between stacked ML models and traditional ML models using 20 acoustic features. The accuracy rates were 97.05% for stacked neural networks (NN), 96.72% for stacked DT, and 96.72% for stacked SVM. In contrast, individual models such as classification and regression trees (CART), NN, and SVM yielded accuracies of 95.10%, 95.52%, and 95.78%, respectively. Sahney *et al.* [13] achieved 98% accuracy using the XGBoost algorithm within a 0-280 Hz frequency range. Raahul *et al.* [14] conducted a comparison of 5 ML algorithms, namely linear discriminant analysis (LDA), KNN, CART, SVM, and RF where SVM outperformed the others in classification. Nashipudimath *et al.* [15] obtained accuracy for gender

recognition and emotion recognition 98.88% and 72.02% respectively using voice activity detection (VAD), MFCC, PCA, and SVM. Kannapiran *et al*. [16] proposed an approach of forward rajan transform (FRT) feature extraction with a light gradient boosting machine (LightGBM) classifier in different datasets like speech accent archive (SAA), common voice (CV), TIMIT Acoustic-Phonetic Continuous Speech Corpus, and voice gender dataset (VGD). VGD achieved an accuracy of 91.8%, 96%, 99.9%, and 93.5%, respectively 0.958.

Some researchers explore feature extraction techniques for speech recognition work and used DL approach. On downsampled speech samples using x- and d-vector algorithms, Kwasny and Hemmerling [17] achieved a 99.60% accuracy in a comparable context by using two-stage transfer learning and the QuartzNet embedder, among other x-vector embedder techniques. Utilizing temporal convolutional networks (TCN) with WavLM pre-trained features, Lebourdais *et al*. [18] got a comprehensive accuracy of 92.1% in gender detection, for male it was 97.8% and for females 94.9%. TCN showed 63.4% accuracy for overlapped speech detection (OSD). Forty male and forty female participants made up the dataset. Ertam [19] developed a reliable technique for gender detection obtaining an incredible 98.4% accuracy in gender prediction by using aural clues to establish gender in a dataset of 3168 samples that were uniformly split between male and female voices. Liztio *et al*. [20] deployed a backpropagation neural network (BPNN) for gender detection utilizing 100 personal records. Discrete Fourier transform (DFT) was utilized to extract features. On a private dataset, the overall accuracy was 72% (76% for males and 68% for females), while on the Kaggle dataset, the overall accuracy was 73% (88% for males and 58% for females).

Many researchers utilize a combination of DL and ML techniques. These papers are examined to provide insight into the integration of multiple methodologies in speech-related research. Sánchez-Hevia *et al.* [21] employed i-vectors and the STFT to classify age and gender. The models utilised, CNN, temporal convolutional network (TCN), and support vector regression, yielded 81% convolutional recurrent neural network (CRNN), 79% convolutional temporal convolutional network (CTCN), 72% TCN, and 76% CNN precision rates. Nugroho *et al*. [22] employed DL to attain outstanding gender recognition accuracy in Javanese speech. Success was aided by approaches such as singular value decomposition (SVD), LR, and SVM also gave good results with accuracies of 95.76%, 93.33%, and 97.78%, respectively. Speech emotion recognition for both genders, including a spectrum of emotions, was the main focus for Mishra and Sharma [23] on a range of datasets named RAVDESS, CREMA-D, SAVEE, and TESS. The CNN+GAP model achieved a stunning 92.28% accuracy rate in detecting emotions from audio data. The evaluation comprises CNN, SVM, and MFCC models. Yu *et al*. [24] used the LibriTTS and VCTK databases for gender-free speech style transfer in a related investigation. The method included a speech gender encoder, text-to-speech (TTS) synthesizer, neural vocoder network, and rule-based model. Livieris *et al*. [25] used a variety of semi-supervised models and algorithms with 366 voice samples in ".wav" format. Similarly, Jayasankar *et al*. [26] collected metrics including zero crossing rate, energy entropy, and short-term energy from 80 voice signals to establish gender with accuracy and precision were 79% and 83%, respectively. To increase privacy in speech recognition, Stoidis and Cavallaro [27] developed gender-ambiguous sounds using the LibriSpeech dataset. GenGAN, a generative adversarial network (GAN), produced voices that were uncertain about gender with an efficacy of 76.64%.

## 3. METHOD

There are numerous steps in the gender recognition process in Bengali, and each one is crucial to the overall. Data collection is the first step, which is then followed by feature extraction, data preparation, and algorithm implementation. Although there are a variety of ways available for accomplishing this work, Figure 1 depicts the strategy we opted to use.

### 3.1. Dataset

Data acquisition is a complex and time-consuming process, particularly when utilized for research; this difficulty arises when speech data in a language like Bengali needs to be acquired. We diligently collected 3185 data points, which included 1100 samples of men, 1035 samples of women, and 1050 samples from persons who identify as third gender. Figure 2 explains a clear outlook of the dataset class distribution. A series of twenty Bengali sentences, each lasting no more than two to three seconds, were provided to each participant. The purpose was to record a wide diversity of language expressions in a limited length of time. Mobile phones were used for all audio recordings, giving a handy and approachable technique for capturing data. This rigorous strategy of acquiring data guarantees the diversity and richness required for a complete analysis of gender recognition in Bengali speech.

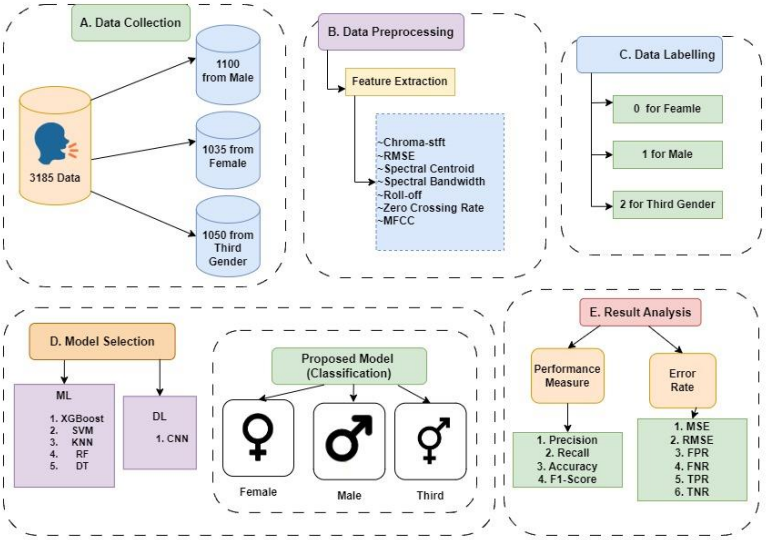


Figure 1. Working method


Figure 2. Dataset class distribution

### 3.2. Feature extraction

Feature extraction is a crucial method in audio processing that extracts and highlights key auditory elements like pitch, spectral content, and temporal aspects. This process streamlines data processing and aids in tasks like speaker identification and speech recognition. It goes beyond ML and pattern detection, enhancing computational efficiency and understanding speech patterns [28], particularly in research on gender recognition in Bengali speech.

#### 3.2.1. Chroma feature

"Chroma" refers to the pitch rotation angle around a helix, revealing energy distribution within specific pitch classes. A Chroma STFT graphic shows energy distribution over time, using the STFT to divide an audio input into frequency components. Figure 3 portrays the chromagram and frequency amplitude for male, female and third gender voice samples.

The time-frequency representation(TFR) technique currently possesses substantial mathematical foundations. Let's assume that the discrete signal $x(n)$ is sampled at a sampling frequency of Fs (sampling frequency) in the time domain. STFT is applied to a signal by multiplying it with a window function $w(n)$ at a specified time instant $t$ [29].

$$X(t,k) = \sum_{n=0}^{M-1} w(n)x\,(n+t)e^{-2\pi jnk/M} \quad (1)$$

#### 3.2.2. Spectral centroid

The middle frequency or average location within a signal's spectrum is expressed by a metric known as the spectral centroid. The weights are determined by averaging the frequencies in the signal and comparing them to the magnitudes of the distinct spectral elements.

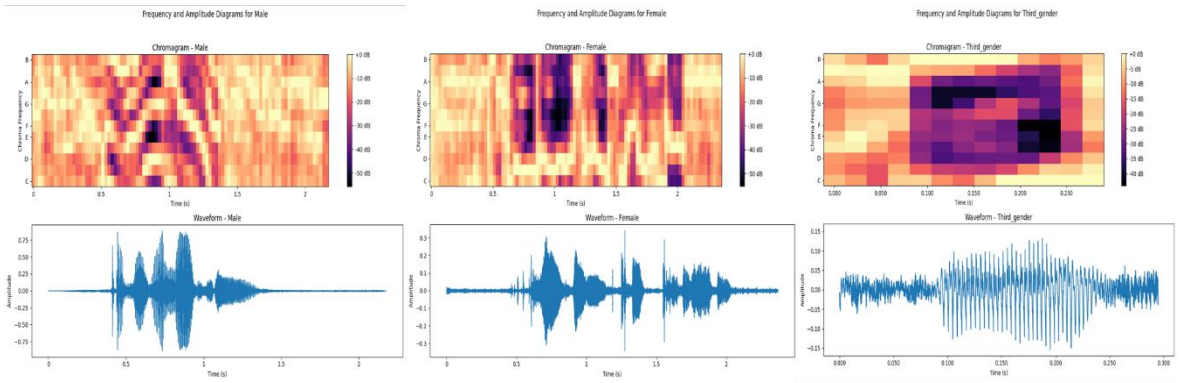


Figure 3. Chromagram and frequency amplitude of male, female, and third gender

### 3.2.3. Spectral bandwidth

A measurement called spectral bandwidth helps to clarify the frequency spectrum of a signal by providing information about the energy distribution across various frequencies as well as the dispersion or concentration of spectral components. Figure 4 has the spectrogram feature of male, female, and third gender audios.

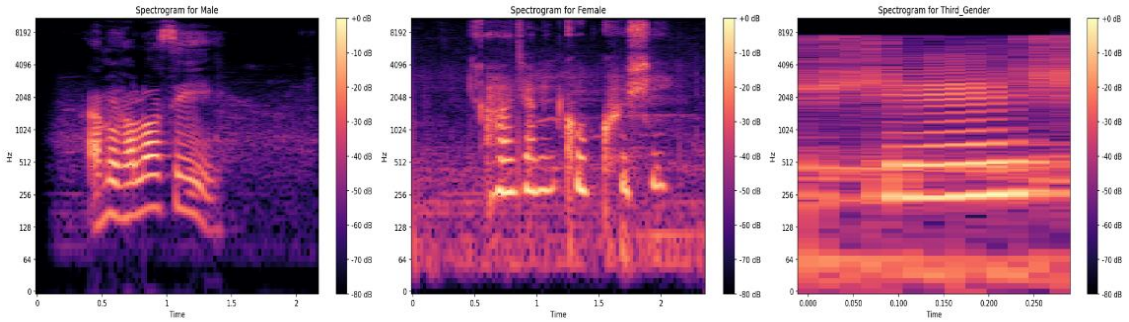


Figure 4. Spectrogram of male, female, and third gender

### 3.2.4. Roll-off

Roll-off offers information about the energy distribution and spectral features of a signal. The definition of this metric is a frequency threshold that encloses a specified percentage of the total spectral energy in a statistic illustration. It mathematically can be represented as (2) [30]:

$$rolloff = \sum_{k=b_1}^{d} s_k = N\left(\sum_{k=b_1}^{b_2} s_k\right) \quad (2)$$

### 3.2.5. Zero crossing rate

The pace at which a signal changes sign inside a frame is known as the zero-crossing rate. Formally defined as (3) [31]:

$$zcr = \frac{1}{T}\sum_{t=0}^{T-1} \| \{s_t s_t - 1 < 0\} \quad (3)$$

Figure 5 contains the boxplot diagram for both roll-off and zero crossing rates of male, female, and third gender.

### 3.2.6. Mel-frequency cepstral coefficients

MFCCs simulate human hearing by using the discrete cosine transform on the logarithm of the signal's power spectrum, produced via the Fourier transform. They record the short-term power spectrum and are commonly used in signal processing, aligning with human hearing features. The MFCC of the male female and third gender data sample is shown in Figure 6. Where x-axis represents time and y-axis represents MFCC coefficients (frequency-related features).

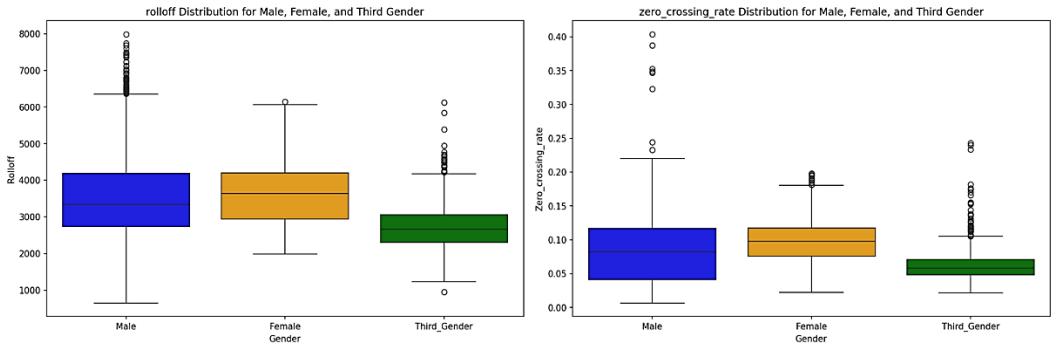


Figure 5. Boxplot for roll-off and zero crossing rate

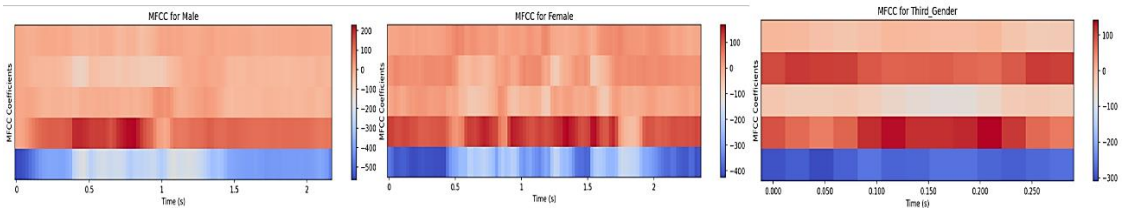


Figure 6. MFCC of male, female, and third gender

DFT to extract information in the frequency domain [32]:

$$X[k] = \sum_{n=0}^{N-1} \times [n] exp\left(-j\frac{2\pi}{N}kn\right) \quad (4)$$

In summary, we compiled a comprehensive set of 26 voice features through our data collection and preparation efforts. These include various characteristics like roll-off, zero crossing rate, spectral centroid, spectral bandwidth, root mean square error, chroma features, and a set of twenty additional features derived from MFCCs. In a CSV file, these organized attributes were stored, facilitating the easy undertaking of algorithms.

### 3.3. Classification techniques

#### 3.3.1. Machine learning approaches

To identify the optimal model, we have explored various machine-learning algorithms. We used a variety of methods, such as DT, KNN, SVM, RF, and XGBoost.

#### 3.3.2. Proposed convolutional neural network approach

A 1D CNN's convolutional layers are crucial for training hierarchical features and assessing one-dimensional sequential input, making it useful for time series and signal processing, including voice-related tasks like gender recognition in Bengali speech. Tables 1 and 2 have the details about the model parameters and the model architecture.

Table 1. Parameter tuning of proposed CNN technique

| Parameters | Epoch | Batch | Optimizer | Activation |
|---|---|---|---|---|
| | 15 | 32 | Adam | Softmax |

Table 2. Proposed CNN model architecture

| Type of layer | Output shape | Parameter |
|---|---|---|
| Conv1D | 43×64 | 256 |
| Dropout | 43×64 | 0 |
| MaxPooling1D | 21×64 | 0 |
| Flatten | 1344 | 0 |
| Dense | 128 | 172160 |
| Dense | 3 | 129 |
| Total params | 172,545 | |
| Trainable params | 172,545 | |
| Non-trainable params | 0 | |

## 4. RESULT DISCUSSION AND ANALYSIS

A spectrum of performance evaluations, including accuracy, precision, recall, F1 score, and different error rates of all algorithms in this part, were investigated. Accompanying the assessment technique were extensive reasons and graphical descriptions in Figure 7. Our findings demonstrated that CNN performed better than other standard machine-learning models [33]. The Figures 8 and 9 contain the accuracy-loss graph and confusion matrix of the proposed CNN approach.

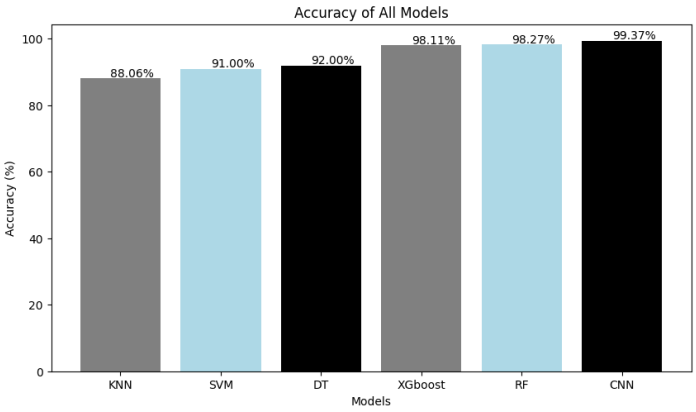


Figure 7. Model accuracy

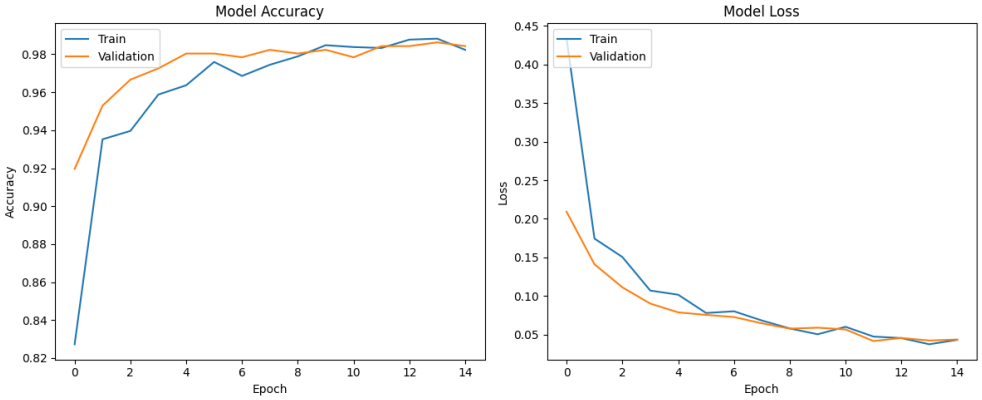


Figure 8. Accuracy and loss graph of CNN

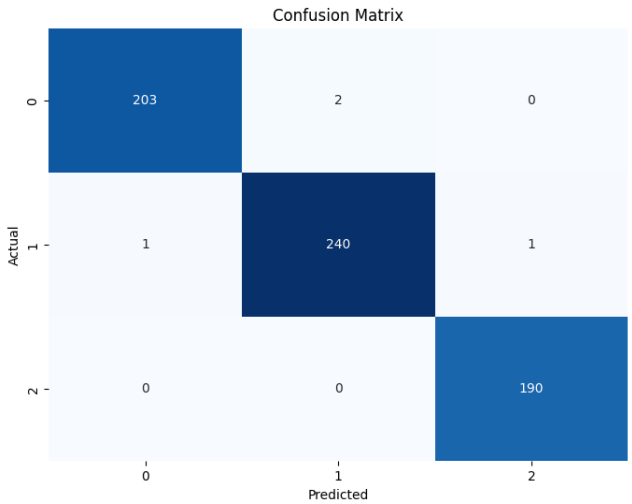


Figure 9. Confusion matrix of CNN

Performance metrics, or scores, are critical for evaluating ML models to evaluate how effective they are at particular tasks. In addition, Table 3 provides the weighted averages of the following metrics: root mean squared error (RMSE), mean squared error (MSE), false positive rate (FPR), false negative rate (FNR), true positive rate (TPR), and true negative rate (TNR); these values are applied to all applied algorithms. These broad metrics offer a nuanced perspective of each model's, advantages and weaknesses, permitting a careful and knowledgeable evaluation of how well each model performs concerning gender detection in Bengali speech. MSE is the square root of RMSE, which measures the intermediate squared difference between expected and actual values. FPR measures the percentage of actual negatives that are mistakenly predicted as positives; on the other hand, FNR measures the percentage of actual positives that are mistakenly predicted as negatives. The percentage of actual positives that are correctly predicted is measured by the TPR, while the percentage of actual negatives that are correctly predicted is assessed by the TNR [34]. Concerning Table 3, the metrics with the lowest values indicate that CNN outperforms the other networks in terms of MSE, RMSE, FPR, and FNR.

Table 3. Performance evaluation with error rate for all models

| Algorithms | Precision | Recall | F1-score | MSE | RMSE | FPR | FNR | TPR | TNR |
|---|---|---|---|---|---|---|---|---|---|
| KNN | 0.88 | 0.88 | 0.88 | 0.166 | 0.407 | 0.095 | 0.152 | 0.847 | 0.904 |
| SVM | 0.91 | 0.91 | 0.91 | 0.098 | 0.314 | 0.121 | 0.133 | 0.866 | 0.878 |
| DT | 0.92 | 0.92 | 0.92 | 0.114 | 0.338 | 0.125 | 0.063 | 0.936 | 0.875 |
| XGBoost | 0.98 | 0.98 | 0.98 | 0.023 | 0.153 | 0.029 | 0.016 | 0.983 | 0.970 |
| RF | 0.98 | 0.98 | 0.98 | 0.020 | 0.142 | 0.019 | 0.016 | 0.983 | 0.980 |
| CNN | 0.99 | 0.99 | 0.99 | 0.006 | 0.079 | 0 | 0.006 | 0.993 | 0 |

## 5. CONCLUSION

In summary, our study effectively developed a method for identifying gender in Bengali speech, producing a distinct and diverse dataset that includes a range of Bengali conversations. The empirical evaluation's conclusions showed how well several algorithms; CNN, XGBoost, SVM, KNN, DTC, and RF predict genders in speech samples from Bengalis with noteworthy accuracy rates ranging from 88% to 99%. This paper addresses gender categorization in non-English languages (Bengali in particular), which closes a significant gap in the literature. It provides useful insights into the performance of several algorithms in terms of gender prediction in this linguistic context. Many lines with various speech tones and emotional expressions were added to the dataset, increasing its richness and helping offset the study's small number of speakers.

Our future research aims to explore gender-based emotion identification and how feelings could vary between genders to understand better the connection between gender and emotional expression in Bengali speech. This work lays the foundation for future research on this topic, which advances our understanding of gender recognition and emotional dynamics in non-English languages.

## REFERENCES


[1] A. Joshi, M. Mahmud, R. G. Ragel, and N. V. Thakur, Eds., *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, vol. 191. in Lecture Notes in Networks and Systems, vol. 191. Singapore: Springer Singapore, 2022, doi: 10.1007/978-981-16-0739-4.

[2] N. M and A. S. Ponraj, "Speech Recognition with Gender Identification and Speaker Diarization," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*, IEEE, Nov. 2020, pp. 1–4, doi: 10.1109/INOCON50539.2020.9298241.

[3] A. A. Alnuaim *et al*., "Speaker Gender Recognition Based on Deep Neural Networks and ResNet50," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1–13, Mar. 2022, doi: 10.1155/2022/4444388.

[4] I. Anwar, "Biology: There is science behind third gender identities." Available [Online]: https://medicine.careers360.com/articles/what-is-the-science-behind-transgender-identities-premium. (Accessed Apr. 05, 2024).

[5] "Front Matter," *Scientific American Mind*, Oct. 22, 2010. Available [Online]: http://www.jstor.org/stable/24943008. (Accessed Apr. 05, 2024).

[6] A. Pahwa and G. Aggarwal, "Speech Feature Extraction for Gender Recognition," *International Journal of Image, Graphics and Signal Processing*, vol. 8, no. 9, pp. 17–25, Sep. 2016, doi: 10.5815/ijigsp.2016.09.03.

[7] S. Safavi, M. Russell, and P. Jančovič, "Automatic speaker, age-group and gender identification from children's speech," *Computer Speech and Language*, vol. 50, pp. 141–156, Jul. 2018, doi: 10.1016/j.csl.2018.01.001.

[8] R. S. Alkhawaldeh, "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network," *Scientific Programming*, vol. 2019, pp. 1–12, Sep. 2019, doi: 10.1155/2019/7213717.

[9] G. Sharma and S. Mala, "Framework for gender recognition using voice," in *Proceedings of the Confluence 2020 - 10th International Conference on Cloud Computing, Data Science and Engineering*, IEEE, Jan. 2020, pp. 32–37, doi: 10.1109/Confluence47617.2020.9058146.

[10] S. Hamdi, A. Moussaoui, M. Oussalah, and M. Saidi, "Gender Identification from Arabic Speech Using Machine Learning," *International Symposium on Modelling and Implementation of Complex Systems*, 2021, pp. 149–162, doi: 10.1007/978-3-030-58861-8_11.

[11] A. Guerrieri, E. Braccili, F. Sgrò, and G. N. Meldolesi, "Gender Identification in a Two-Level Hierarchical Speech Emotion Recognition System for an Italian Social Robot," *Sensors*, vol. 22, no. 5, p. 1714, Feb. 2022, doi: 10.3390/s22051714.

[12] P. Gupta, S. Goel, and A. Purwar, "A Stacked Technique for Gender Recognition Through Voice," in *2018 11th International*

*Conference on Contemporary Computing, IC3 2018*, IEEE, Aug. 2018, pp. 1–3, doi: 10.1109/IC3.2018.8530520.
[13] S. Sahney, T. Babu, K. Narahari, and A. Karan, "Identifying the gender of a voice using acoustic properties," *International Journal of Advances in Engineering Research (IJAER)*, vol. 20, no. 6, pp. 13–20, 2020.
[14] A. Raahul, R. Sapthagiri, K. Pankaj, and V. Vijayarajan, "Voice based gender classification using machine learning," *IOP Conference Series: Materials Science and Engineering*, vol. 263, no. 4, p. 042083, Nov. 2017, doi: 10.1088/1757-899X/263/4/042083.
[15] M. M. Nashipudimath, P. Pillai, A. Subramanian, V. Nair, and S. Khalife, "Voice Feature Extraction for Gender and Emotion Recognition," *ITM Web of Conferences*, vol. 40, p. 03008, Aug. 2021, doi: 10.1051/itmconf/20214003008.
[16] P. Kannapiran and M. M. R. Sindha, "Voice-Based Gender Recognition Model Using FRT and Light GBM," *Tehnicki Vjesnik*, vol. 30, no. 1, pp. 282–291, Feb. 2023, doi: 10.17559/TV-20220302182704.
[17] D. Kwasny and D. Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors*, vol. 21, no. 14, p. 4785, Jul. 2021, doi: 10.3390/s21144785.
[18] M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier, "Overlapped speech and gender detection with WavLM pre-trained features," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2022-September, pp. 5010–5014, Sep. 2022, doi: 10.21437/Interspeech.2022-10825.
[19] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Applied Acoustics*, vol. 156, pp. 351–358, Dec. 2019, doi: 10.1016/j.apacoust.2019.07.033.
[20] L. M. Liztio, C. A. Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto, "Gender identification based on speech recognition using backpropagation neural network," in *Proceedings - 2020 International Seminar on Application for Technology of Information and Communication: IT Challenges for Sustainability, Scalability, and Security in the Age of Digital Disruption, iSemantic 2020*, IEEE, Sep. 2020, pp. 88–92, doi: 10.1109/iSemantic50169.2020.9234237.
[21] 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, Jan. 2022, doi: 10.1007/s11042-021-11614-4.
[22] K. Nugroho, E. Noersasongko, Purwanto, Muljono, and H. A. Santoso, "Javanese Gender Speech Recognition Using Deep Learning and Singular Value Decomposition," in *Proceedings - 2019 International Seminar on Application for Technology of Information and Communication: Industry 4.0: Retrospect, Prospect, and Challenges, iSemantic 2019*, IEEE, Sep. 2019, pp. 251–254, doi: 10.1109/ISEMANTIC.2019.8884267.
[23] P. Mishra and R. Sharma, "Gender Differentiated Convolutional Neural Networks for Speech Emotion Recognition," in *International Congress on Ultra Modern Telecommunications and Control Systems and Workshops*, IEEE, Oct. 2020, pp. 142–148, doi: 10.1109/ICUMT51630.2020.9222412.
[24] C. Yu, C. Fu, R. Chen, and A. Tapus, "First Attempt of Gender-free Speech Style Transfer for Genderless Robot," in *ACM/IEEE International Conference on Human-Robot Interaction*, IEEE, Mar. 2022, pp. 1110–1113, doi: 10.1109/HRI53351.2022.9889533.
[25] I. E. Livieris, E. Pintelas, and P. Pintelas, "Gender Recognition by Voice Using an Improved Self-Labeled Algorithm," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 492–503, Mar. 2019, doi: 10.3390/make1010030.
[26] T. Jayasankar, K. Vinothkumar, and A. Vijayaselvi, "Automatic gender identification in speech recognition by genetic algorithm," *Applied Mathematics and Information Sciences*, vol. 11, no. 3, pp. 907–913, May 2017, doi: 10.18576/amis/110331.
[27] D. Stoidis and A. Cavallaro, "Generating gender-ambiguous voices for privacy-preserving speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, ISCA: ISCA, Sep. 2022, pp. 4237–4241, doi: 10.21437/Interspeech.2022-11322.
[28] R. Hibare and A. Vibhute, "Feature Extraction Techniques in Speech Processing: A Survey," *International Journal of Computer Applications*, vol. 107, no. 5, pp. 1–8, 2014, doi: 10.5120/18744-9997.
[29] N. Kehtarnavaz, "Frequency Domain Processing," in *Digital Signal Processing System Design*, Elsevier, 2008, pp. 175–196, doi: 10.1016/b978-0-12-374490-6.00007-6.
[30] R. Ramnauth, "Algorithmic Audio Feature Extraction in English," 2020. Available [Online]: https://rramnauth2220.github.io/blog/posts/code/200525-feature-extraction.html. (Accessed Apr. 05, 2024).
[31] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/Unvoiced decision for speech signals based on Zero-Crossing Rate and energy," *Advanced Techniques in Computing Sciences and Software Engineering (conference)*, 2009, pp. 279–282, doi: 10.1007/978-90-481-3660-5_47.
[32] S. R. Kulkarni, "Frequency domain and fourier transforms," *Information Sciences and Systems, Princeton University*, 2002.
[33] F. Akdeniz and Y. Becerikli, "Performance Comparison of Support Vector Machine, K-Nearest-Neighbor, Artificial Neural Networks, and Recurrent Neural networks in Gender Recognition from Voice Signals," *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, 2019, pp. 1-4, doi: 10.1109/ISMSIT.2019.8932818.
[34] M. M. Nasef, A. M. Sauber, and M. M. Nabil, "Voice gender recognition under unconstrained environments using self-attention," *Applied Acoustics*, vol. 175, p. 107823, Apr. 2021, doi: 10.1016/j.apacoust.2020.107823.

## BIOGRAPHIES OF AUTHORS

**Habiba Dewan Arpita** is currently pursuing her undergraduate studies in Computer Science and Engineering at Daffodil International University. Her research interests revolve around ML, deep learning, NLP, signal processing, and computer vision. She actively contributes to ongoing research in these domains. She can be contacted at email: habiba15-14042@diu.edu.bd.

**Abdullah Al Ryan** completed his B.Sc. in Computer Science and Engineering from Daffodil International University, Bangladesh, in 2023. Later, he worked as an Jr. Software Engineer (AI/ML) at Nexalinx. Currently, he is pursuing Joint Master's degree in Computational Color and Spectral Imaging (COSI). His research interests encompass data science, machine learning, deep learning, computer vision, natural language processing, and image processing. As an active researcher, he frequently reviews papers for conferences. He can be contacted at email: abdullah15-13088@diu.edu.bd.

**Md. Fahad Hossain** completed his Bachelor of Science (B.Sc.) and Master of Science (M.Sc.) degrees in Computer Science and Engineering at Daffodil International University, achieving top academic honors. He currently serves as a Senior Lecturer in the Department of Computer Science and Engineering at Bangladesh University. With numerous publications in reputable international and national journals and conference proceedings, his research interests cover diverse fields, including natural language processing, multi-label offensive text detection in Bangla, sentiment analysis, epidemiological modeling, and ethical considerations in data-driven research. He can be contacted at email: fahad.hossain@bu.edu.bd.

**Md. Sadekur Rahman** completed his Bachelor of Science (B.Sc.) and Master of Science (M.Sc.) degrees in Applied Mathematics and Informatics at the Peoples' Friendship University of Russia. Currently, he holds the position of Assistant Professor at the Department of Computer Science and Engineering at Daffodil International University. Additionally, he serves as an advisor to the DIU NLP and ML Research Lab. With numerous publications in both international and national journals and conference proceedings, his research interests span areas such as data mining, artificial intelligence, pattern recognition, and natural language processing. He can be contacted at email: sadekur.cse@daffodilvarsity.edu.bd.

**Md Sajjad** is a cybersecurity expert with specialized knowledge in machine learning, artificial intelligence, and big data analytics for advanced security solutions. Currently studying at Western Sydney University, Australia, he is pursuing a Master's in Information and Communication Technology with a focus on Computer, Network Security, Data, and Math Sciences. Sajjad is dedicated to developing intelligent, data-driven methods to safeguard digital infrastructures. He holds a Bachelor's degree in Computer Science and Engineering from American International University-Bangladesh, where he built a solid foundation in cybersecurity frameworks, threat analysis, and predictive defense strategies. Sajjad is committed to staying at the forefront of industry advancements, driven by a passion for cybersecurity resilience and proactive risk management. He can be contacted at email: sajjadmdsaad@gmail.com.

**Nuzhat Noor Islam Prova** is a Data Scientist with expertise in machine learning, artificial intelligence, and statistical analysis. She holds an MS in Data Science from Pace University's Seidenberg School of Computer Science and Information Systems and a Bachelor's degree in Business Administration from North South University. Her research interests span healthcare analytics, computer vision, and natural language processing, with particular focus on explainable AI and transfer learning applications. She is an active member of several professional organizations, including IEEE, ASA, INFORMS, and ADaSci, and holds multiple professional certifications in data science and AI engineering from IBM and Microsoft. She can be contacted at email: nuzhatnsu@gmail.com.