International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Efficacy of Artificial Intelligence for gender classification in speech signals

Olumide Olayode Ajayi [1,*], Ayo Isaac Oyedeji [2], Olusogo Julius Adetunji [2] and Janet Olubunmi Jooda [3]

[1] *Department of Electrical and Electronics Engineering, Faculty of Engineering, Adeleke University, Ede, Nigeria.*
[2] *Department of Computer Engineering, Faculty of Engineering, Olabisi Onabanjo University, Ago Iwoye, Nigeria.*
[3] *Department of Computer Engineering, Faculty of Engineering, Redeemer's University, Ede, Nigeria.*

International Journal of Science and Research Archive, 2025, 16(01), 1818-1829

Publication history: Received on 12 June 2025; revised on 24 July 2025; accepted on 26 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2238

Abstract

The classification or recognition of human voices are required for different applications such as speech emotion recognition, medicals, communications software, and security. However, gender classification of speech signals is complex aspect of speech recognition systems; thereby requiring robust signal processing strategies. This paper investigates the efficacy of Artificial Intelligence (AI) for gender classification in speech signals. Two different AI-based gender voice classifiers namely K-Nearest Neighbor (K-NN) and Long Short-Term Memory (LSTM) were developed. First, speech signals were recorded from different male and female speakers at a sampling rate of 48 kHz. Each of the raw speech signals was filtered and the useful portion of the signal was segmented. The Mel Frequency Cepstral Coefficient (MFCC) and Mel Spectrum (MS) features were extracted from each signal via framing, hamming window, and FFT. An equal number of observations each for the male and female classes were generated. The total 2000 observations were partitioned into 80% for training and 20% for testing. The training dataset was used to train both the K-NN and LSTM classifiers. The results obtained from testing with the testing dataset showed that the K-NN classifier gave precision, recall, accuracy and F1-score values of 0.9852, 0.9850, 0.9925 and 0.9851, respectively, whereas the LSTM classifier gave 0.9132, 0.9091, 0.9525, and 0.9091, respectively. The classifiers achieve more than 0.95 (or 95%) classification accuracy; thereby demonstrating the efficacy of the AI strategies in distinguishing between a male voice and a female voice.

Keywords: Artificial Intelligence (AI); Speech Signal; Gender Voice; Deep Learning (Dl); Machine Learning (Ml)

1. Introduction

Voice classification involves the processing of speech signals. Speech communication is the oldest and most natural method of communication among human beings [a]. Research as shown that there is a difference between the speech acoustic features of the female speakers and male speakers; and the voice of a male human being is distinct from that of a female human by the tune and signal energy [2],[3],[4]. Some areas of application of speech signal processing are speech emotion recognition, medicals, communications software, security and investigation. For instance, the police can use the voice classification system to obtain facts about the speaker in a recorded speech to aid or speed-up their investigations [5]. The types of speech in speech recognition systems are connected word, isolated speech, continuous speech and spontaneous speech [6].

Furthermore, speech recognition (or detection) is an important component of the Human Computer Interface (HCI) technologies such as voice bots and call centers [6],[7]. Automatic Speech Recognition (ASR) systems are faced with some challenges such as background noise, breathy voice, variation in speakers, intonation, stress and deception in speech [5],[8],[9],[10]. Studies have shown that an ASR system can be enhanced by gender-dependent models [11]. In consequence, the development of a gender voice classifier of an acceptable classification accuracy cannot be over-

* Corresponding author: Olumide Olayode Ajayi

emphasized. Thus, this paper investigates the efficacy of Artificial Intelligence (AI) for gender classification in speech signals. Comparisons were made between two AI techniques namely k-Nearest Neighbor (K-NN) and Long Short-Term Memory (LSTM). The K-NN is a Machine Learning (ML) technique [12], whereas LSTM is a Deep Learning (DL) or Deep Neural Network (DNN) technique [13]. ML is a branch of AI and DL is a special aspect of ML that extends the functionality of ML to include tasks such as knowledge representation and feature extraction. The K-NN has found application in pattern recognition whereas LSTM has found application in time-series prediction [14],[15].

The second section of this paper reviews related works in speech/voice classification. The third section discusses the development of the K-NN and LSTM classifiers. The fourth section presents the performance evaluation of the classifiers, and finally the conclusion is discussed in the fifth section.

2. Related work

Some studies have been carried out on the classification or recognition of human voices for different applications. The work in [16] employed linear predictive coding and adaptive neuro-fuzzy for speech recognition in robots. A semi-supervised ensemble model for gender voice classification was presented in [17]. Alcalde (2019) [2] carried out performance evaluations of some ML techniques for gender voice classification in terms of precision and recall. The results revealed that the Support Vector Machine (SVM) technique with sequential minimal optimization outperforms the other techniques. Shaqra et al (2019) [18] developed multilayer perceptron Neural Network (NN) classifiers for improving gender and emotion classification in speech signals. The proposed technique gave higher recognition accuracy compared to a conventional approach of all-in-one model.

A modified Mel Frequency Cepstral Coefficients (MFCC)-based technique was proposed for Bangla language speech recognition system in [19]. Kwasny & Hemmerling (2021) [7] employed DNN for the classification of gender speakers from their signals. Shareef et al. (2020) [20] proposed a decision tree algorithm based on Vector Quantization (VQ) for gender voice recognition. Uddin et al. (2021) [21] applied a multi-layer feature extraction technique with 1D-CNN for gender voice classification improvement.

3. Material and methods

The section presents the methodology for developing the proposed AI-based gender voice classifiers.

3.1. Signal Preprocessing

Figure 1 shows the framework for generating voice signal features. The raw speech signal was filtered using 10th-order low-pass FIR filter to remove noisy background from the signal waveform. Then, the filtered signal waveform was passed through a thresholding to detect the voice region. The detected voice signal was segmented into samples of equal length. Segmenting the voice signal into frames of short length is important in that it helps to overcome variation on a long-time scale. The number of voice samples, n, in a frame is obtained as [2]

$$n \approx K \times F_s \qquad \text{.....(1)}$$

- where K is a standard frame length of 25 ms, and F_s is the sampling rate (Hz)
- The Hamming window's coefficients, w(n), are calculated as [2]

$$w(n) = 0.54 - 4.4\cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \le n \le N-1 \qquad \text{..... (2)}$$

where N is the total number of points (or length) of the Hamming window. The Hamming window helps to minimize signal discontinuities.

3.2. Feature Extraction

In this experiment, two different types of features were extracted from the speech signals. The features are Mel Spectrum and Mel Frequency Cepstral Coefficients (MFCC). The Mel-Spectrogram represents the acoustic time-frequency representation of sound, and the Mel spectrum is obtained as [22]:

$$mel\_spectrum = 2595^{10}\log\left(1 + \frac{f}{700}\right) \qquad \text{..... (3)}$$

where f is the frequency on the mel frequency scale. The coefficients of power spectrogram are used in frequency scale. The MFCC reveals the vocal tract (shape of a human voice), which can be represented by cepstral

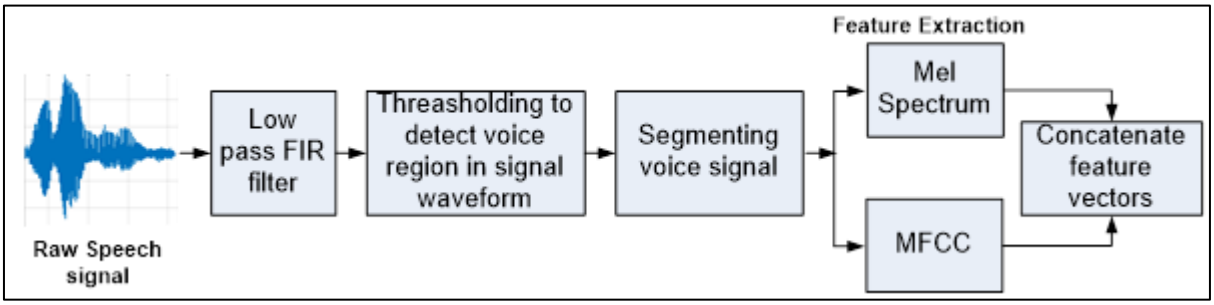

Figure 1 Framework for generating voice signal features

coefficients [21]. The MFCC technique is expatiated in [20]. The procedure for obtaining MFCC is presented in Algorithm 1. The extracted features per gender class were concatenated into a vector as an observation. All the observations in the dataset are of equal feature length (i.e. 280). Each observation was labelled according to its gender class; that is, as "Male" or "Female". In order to make the MFCC and Mel spectrum to be on the same scale, the whole features matrix X was normalized. The normalized features matrix X_{norm} is obtained as:

$$X_{norm} = \frac{(X-\mu)}{\sigma} \quad \text{..... (4)}$$

- where μ is the mean of the features matrix X, and σ is the standard deviation of X.

The matrix X_{norm} is of the same dimension as X; that is, $M \times l$, where M is the number of observations in the matrix and l is the feature length. The classes labels, Y, of the observations is one-dimensional i.e. $M \times 1$. The training of both the K-NN and LSTM are based on supervised learning [12] in which each observation is given a target label during training.

Algorithm 1: Extraction of MFCC from a voice signal

Input: Speech signal

Output: MFCC

BEGIN

Read speech signal

Apply Hamming windowing

Perform Fast Fourier Transform (FFT)

Pass waveform from filter banks

Sum the outputs of filter banks

Perform rectification

Apply Discrete Cosine Transform (DCT) to produce the MFCC

END

3.3. Dataset Description

Speech signals were obtained from 32 male speakers and 23 female speakers. The speakers were between the ages of 18 and 50. All the speakers were made to utter the same sentence ("The world is a global village."). Each speech was recorded with the same mobile phone at a sampling rate of 48 kHz. Each of the gender classes (i.e. male and female) consists of 1000 voice signal observations to make a total of 2000 observations. Each observation consists of 280 features out of which 196 are Mel spectrum features and 84 are MFCC features. Thus, the total features' matrix was of size 2000 x 280. The 2000 observations consisting of both male and female voices were partitioned into 80% for training

dataset and 20% for testing dataset. Each gender class had equal proportion in both the training dataset and testing dataset.

3.4. Training the K-Nearest Neighbor (K-NN) Classifier

The "k" in K-NN is an hyperparameter that denotes the number of observations (greater than one) having mutual similarity in a given instance space, and take part in voting [12, 15]. In other words, given a test observation x, all the observations with the closest similarity with x form its k nearest neighbours. The similarity is determined from distance measures such as the Euclidean distance.

3.5. Training the Long Short-Term Memory (LSTM) Classifier

In practice, the LSTM network is made up of more than one memory block; and each block consists of one or more memory cells. The information to be kept will proceed to the next cell while irrelevant information will be squashed at the forget gate using the sigmoid (sigma) and hyperbolic tangent (tanh) transfer functions. The hidden layer in the LSTM network is a recurrent layer that feeds the input units, input gate, forget gate and output gate from the output units of the cell as shown in Figure 2.

Figure 3 presents the procedure for training the LSTM classifier using the MFCC and Mel Spectrum. The feature vectors were converted to cell arrays which are the suitable forms for training an LSTM network [14].

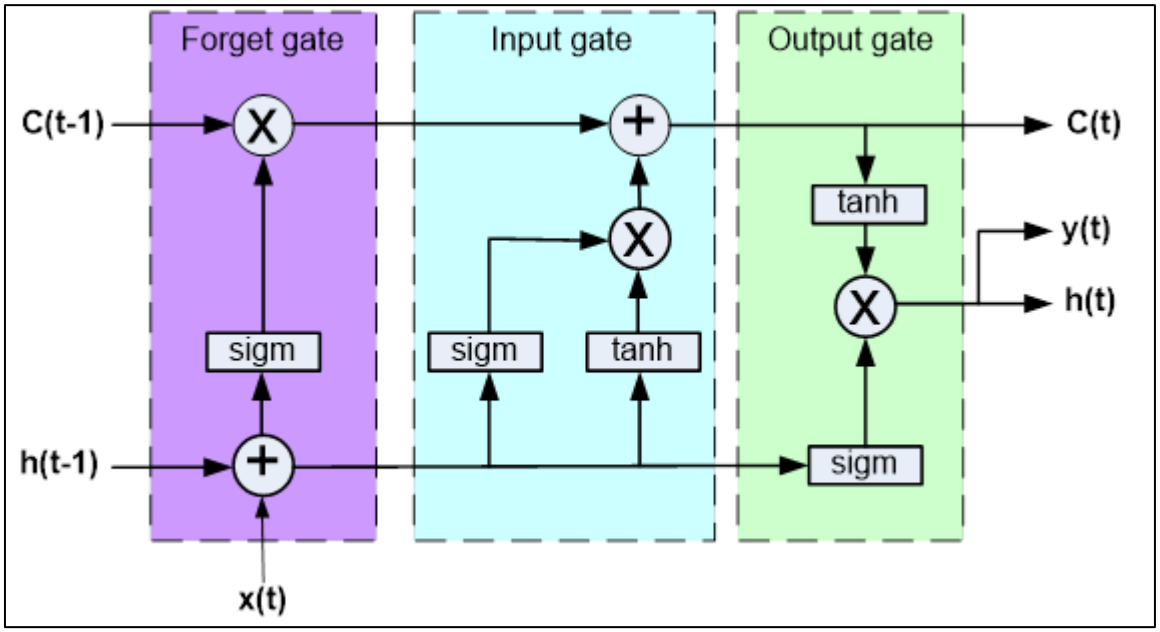

Figure 2 Structure of an LSTM cell [13]

3.6. Performance Evaluation of the Classifiers

Flow chart for the development of the AI-based gender voice classification is shown in Figure 3. The first stage of the process involves loading the features matrix. Then the features matrix is divided into training and testing datasets. For the LSTM, the datasets are converted to cell array. Both the K-NN and LSTM classifiers were trained separately by the features and target gender labels. Finally, the trained classifier was saved for the testing stage.

The metrics used in this study to evaluate the performance of the gender voice classifiers are precision, recall, F1-score and accuracy (of classification). In-depth explanations of these metrics can be found in [12], and are briefly

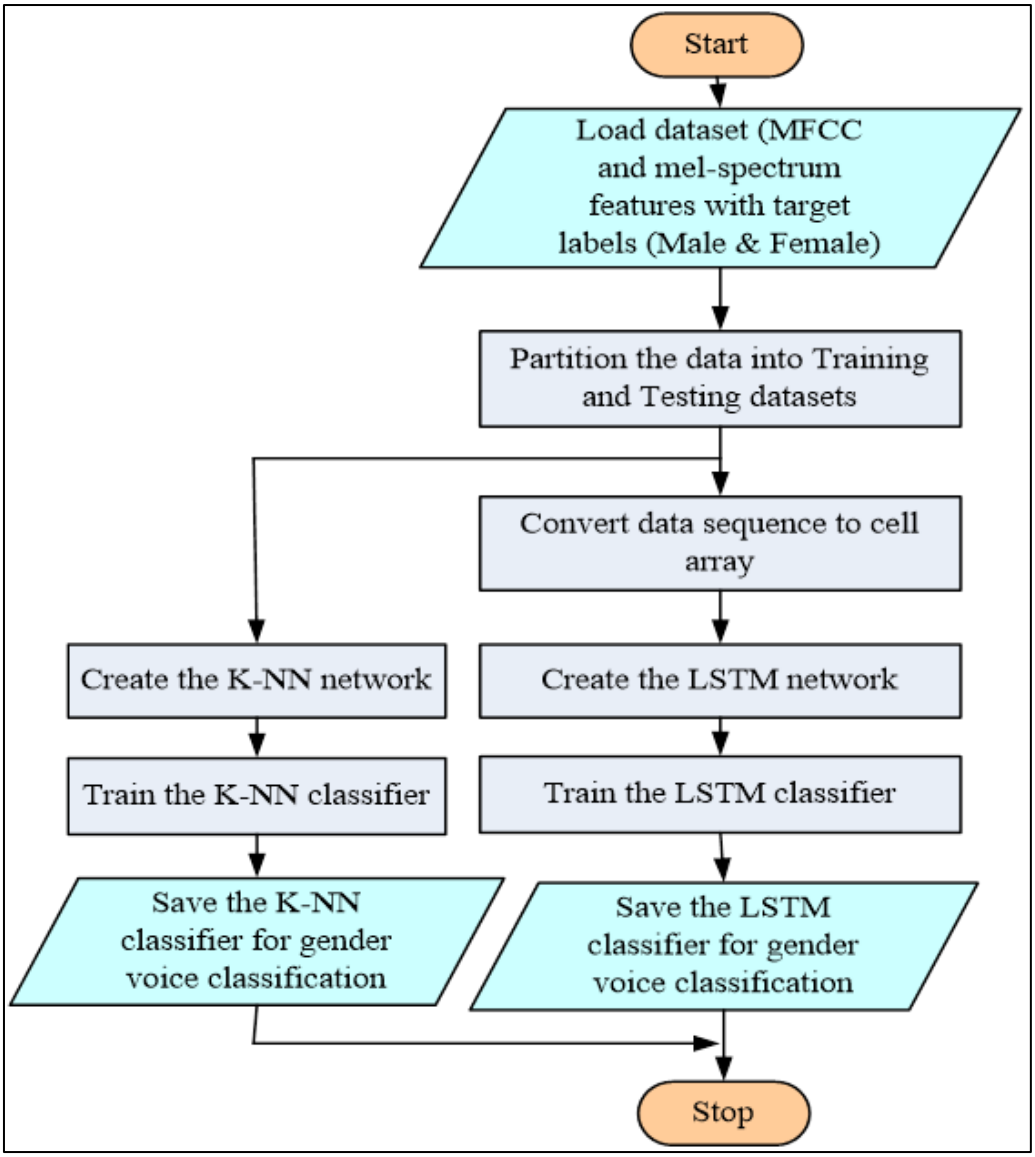

Figure 3 Flow chart for the development of the AI-based gender voice classification

3.6.1. Be scribed as follows

- *Precision* — Precision denotes the probability that the classifier correctly gives a true positive.
- *Recall* — Recall as to do with the number of times that the classifier classifies a particular observation as true positive among all positive observations.
- *F1 score* — F1 score is a measure of the combination of precision and recall by a single quantity.
- *Accuracy* — The classification accuracy is the rate of correctly classified observations, and is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \qquad \ldots\ldots\ldots\ldots \qquad (5)$$

were

- TP (true positive) is when the classifier correctly classifies a positive observation as positive,
- TN (true negative) is when the classifier correctly classifies a negative observation as negative,
- FN (false negative) is when the classifier wrongly classifies a positive observation as negative, and
- FP (false positive) is when the classifier wrongly classifies a negative observation as positive.

The parameters and values used in the simulation of the systems are contained in Table 1.

Table 1 System specifications

Parameter	Value
Sampling Frequency	4.8 kHz
Hamming window duration	0.025 s
Percentage Overlap	35 %
Lowpass Filter	Finite Impulse Response
Number of nearest neighbours in K-NN	3
Distance measure	Euclidean
LSTM Optimizer	adaptive moment estimation (Adam)
Maximum Epochs	200
Batch Size	100
Initial Learning Rate	0.01
Sequence Length	280

4. Results and Discussion

Figures 4 and 5 present the comparisons between the waveforms of raw speech signal and the filtered speech signal for a female speaker and a male speaker, respectively. It is observed that the raw signal has higher amplitudes than the filtered signal (the clean version) due to the noise components contained in the raw signal. However, the low-pass filter eliminates the noise components to produce a clean version of the signal suitable for training a classifier.

The confusion matrix for the K-NN and LSTM classifiers for gender voice classification are presented in Figure 6 and Figure 7, respectively. Out of the 400 voice samples in the testing dataset, the K-NN classifier gave 200 true positives (TP), 0 false negatives (FN), 197 true negatives (TN) and 3 false positives (FP). On the other

hand, the LSTM classifier gave 200 TP, 0 FN, 181 TN, and 19 FP. The results reveal that both the K-NN and LSTM classifiers correctly classified all the male voices but misclassified 3 and 19 female voices, respectively.

The precision, recall, F1-score and accuracy performances of the gender voice classifiers developed in this study are presented in Figures 8, 9, 10 and 11, respectively. The K-NN and LSTM classifiers achieved precision values 0.9852 and 0.9132, respectively. The recall values 0.9850 and 0.9091 were achieved for K-NN and LSTM, respectively. The K-NN classifier achieved F1-score of 0.9851 whereas the LSTM achieved 0.9091. For the classification accuracy, the K-NN classifier gave 0.9925 whereas the LSTM gave 0.9091.

The results reveal excellent performances by both classifiers as they gave over 90 % in precision, recall, F1-score and accuracy. Furthermore, the K-NN classifier gave relatively higher precision, recall, F1-score and accuracy values compared to the LSTM classifier, which is an indication that the K-NN classifier is able to discriminate between the male and female features better than the LSTM classifier.

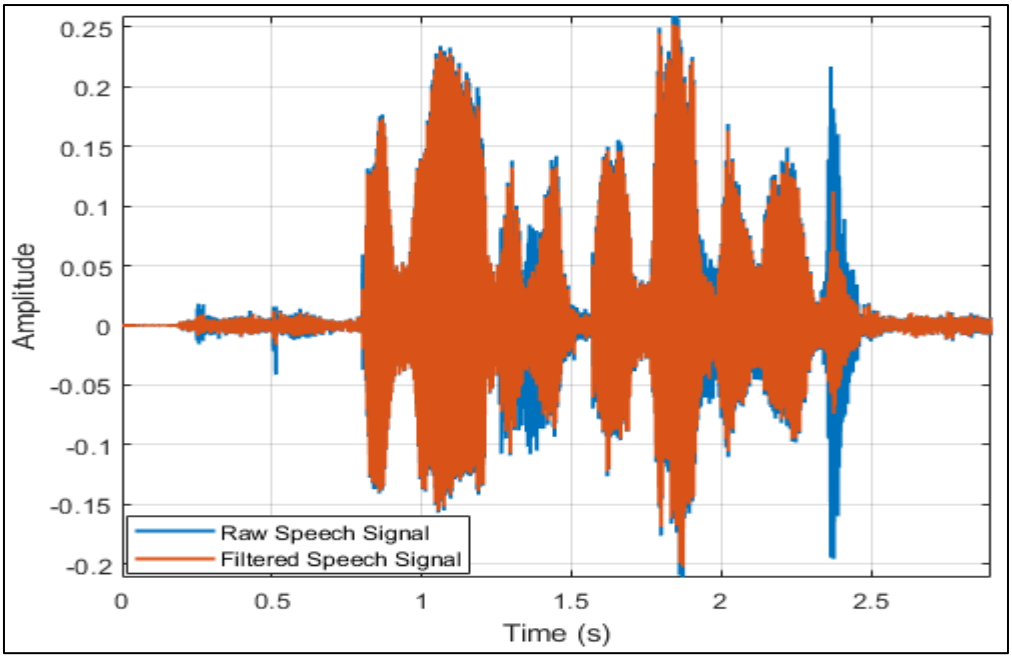

Figure 4 Speech signal waveform of a female speaker (AUD-20211126-WA0006.aac)

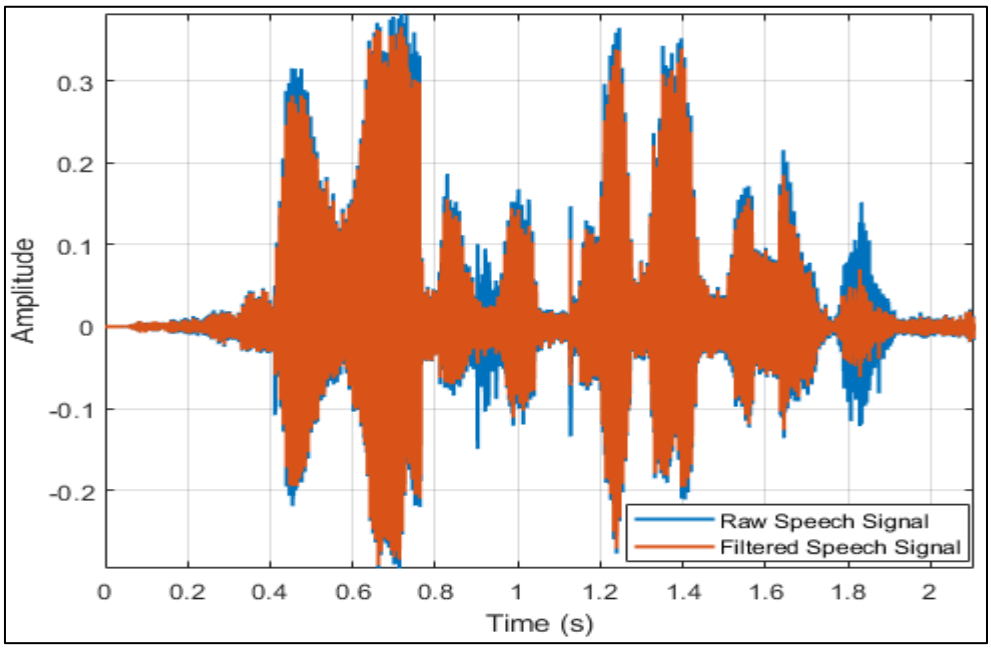

Figure 5 Speech signal waveform of a male speaker (AUD-20211126-WA0005.aac)

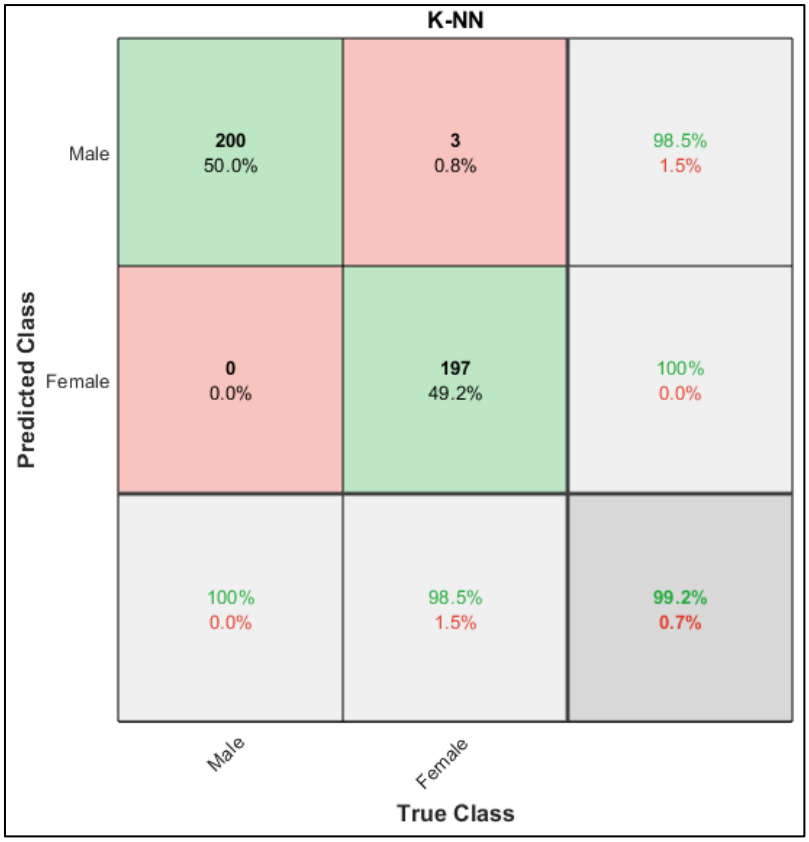

Figure 6 Confusion matrix for gender voice classification by the K-NN classifier

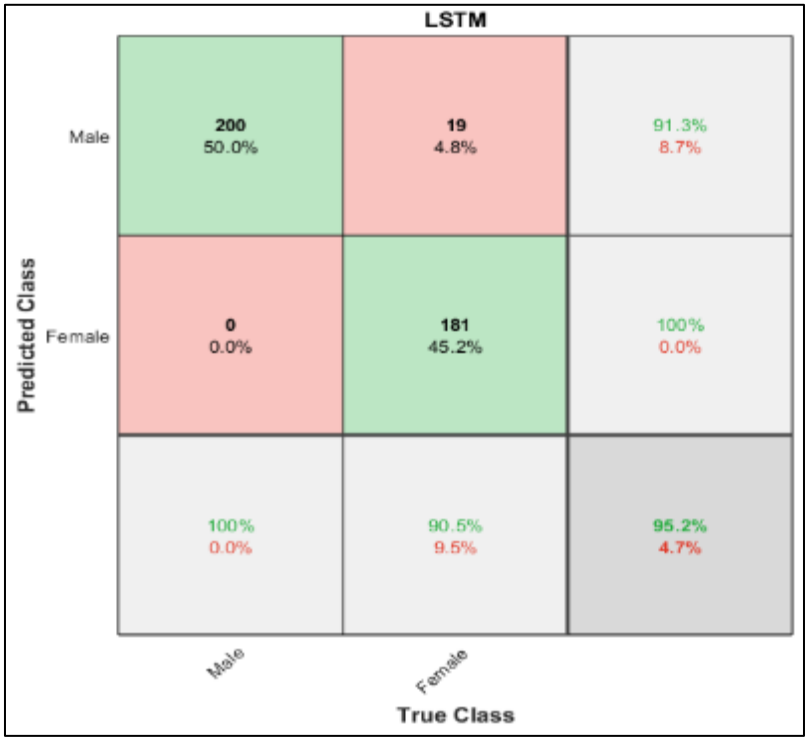

Figure 7 Confusion matrix for gender voice classification by the LSTM classifier

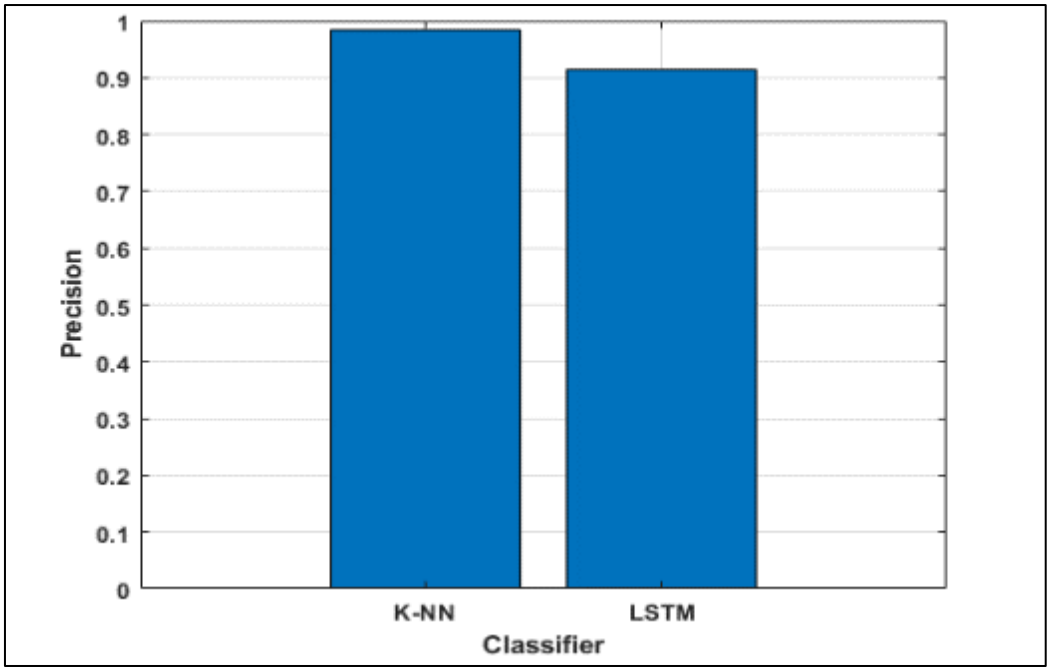

Figure 8 Precision performance of the gender voice classifiers

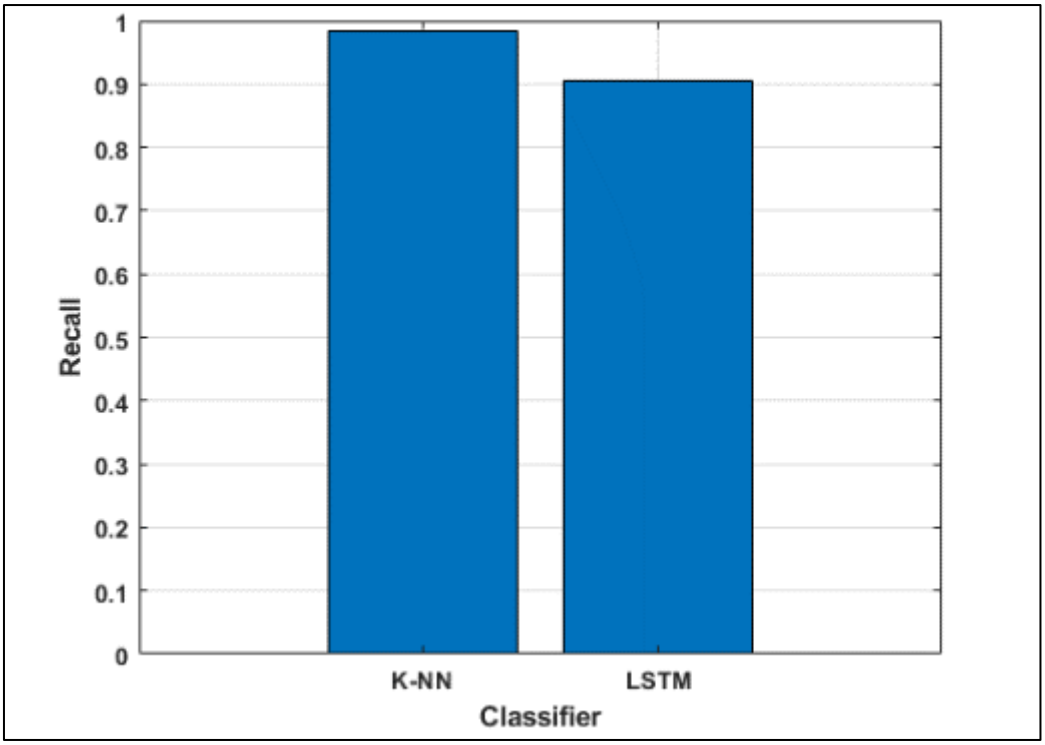

Figure 9 Recall performance of the gender voice classifiers

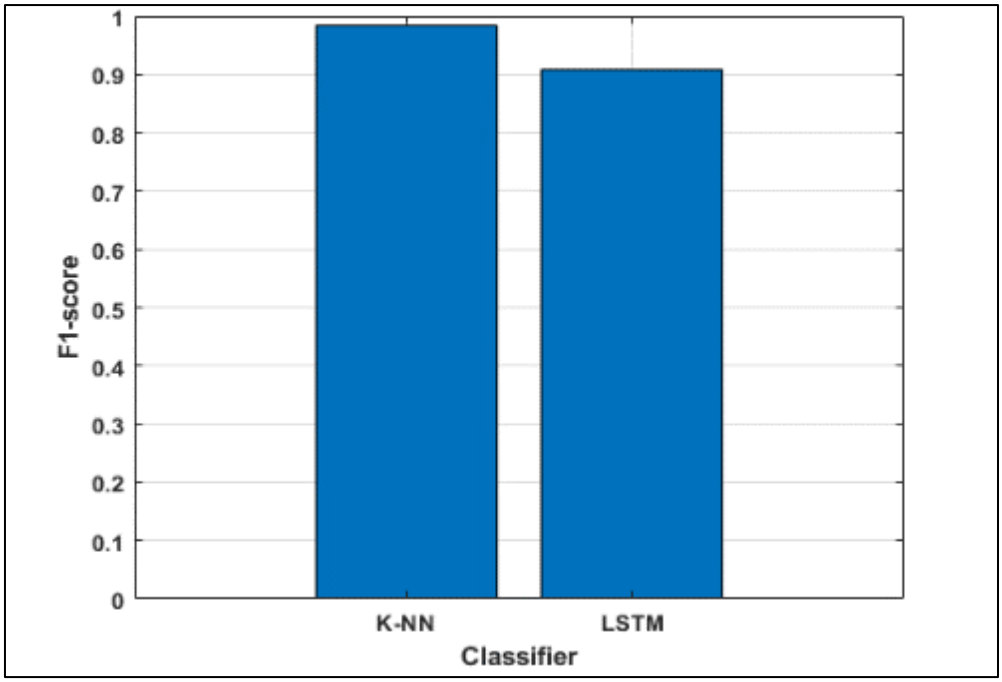

Figure 10 F1-score performance of the gender voice classifiers

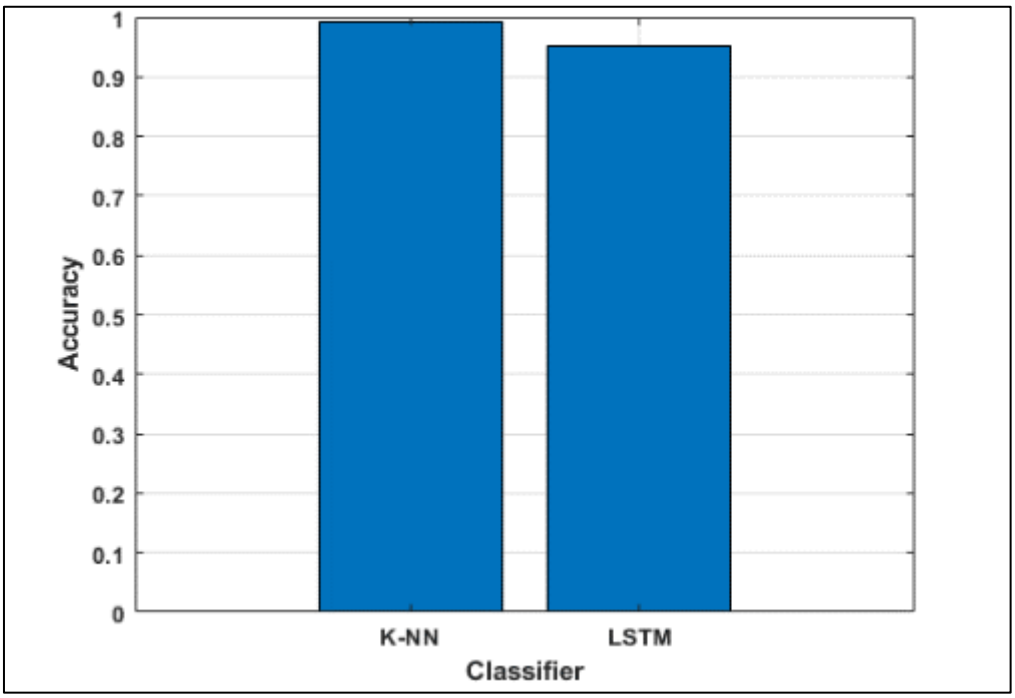

Figure 11 Accuracy performance of the gender voice classifiers

5. Conclusion

This paper investigates the efficacy of Artificial Intelligence (AI) for gender classification in speech signals. Two different AI-based gender voice classifiers namely K-Nearest Neighbor (K-NN) and Long Short-Term Memory (LSTM) were developed. Both the K-NN and LSTM classifiers were trained with different male and female voice signal features. The performance evaluation showed that both classifiers achieve more than 95% classification accuracy; thereby demonstrating the efficacy of the AI strategies in distinguishing between a male voice and a female voice.

Future work will experiment with more feature types in addition to Mel spectrum and MFCC features. Furthermore, more AI techniques would be compared for the same task and the proposed classifiers will be evaluated on different standard datasets.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] O'Shaughnessy DO. Review of methods for coding of speech signals. EURASIP Journal on Audio, Speech, and Music Processing. 2023; 8, 1–25.

[2] Alkhawaldeh RS. DGR:Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network. Hindawi Scientific Programming. 2019; 1-12.

[3] Klatt DH, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 1990; 87(2), 820-857.

[4] Taspinar YS, Saritas MM, Cinar I, Koklu M. Gender Determination Using Voice Data. International Journal of Applied Mathematics, Electronics and Computers. 2020; 8(4), 232-235.

[5] Hollien H, Bahr RH, Harnsberger JD. Issues in Forensic Voice. Journal of Voice. 2013; 28(2), 170–184.

[6] Desai N, Dhameliya PK, Desai PV. Feature Extraction and Classification Techniques for Speech Recognition:A Review. International Journal of Emerging Technology and Advanced Engineering. 2013; 3(12), 1–5.

[7] Kwasny D, Hemmerling D. Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks. Sensors. 2021; 21(4785), 1-18.

[8] Lavan N, Burton AM, Scott SK, Mcgettigan C. Flexible voices:Identity perception from variable vocal signals. Psychonomic Bulletin & Review. 2018; 1-13.

[9] Njie S, Lavan N, Mcgettigan C. Talker and accent familiarity yield advantages for voice identity perception:A voice sorting study. Memory & Cognition. 2023; 2023(51), 175–187.

[10] Patel S, Shrivastav R, Eddins DA. Developing a Single Comparison Stimulus. Journal of Speech, Language, and Hearing Research. 2012; 55(4), 639–648.

[11] Favre B, Grishman R. Speech segmentation and spoken document processing. IEEE Signal Processing Magazine. 2008; 25(3), 59-69.

[12] Kubat M. An Introduction to Machine Learning, 2nd Ed, Springer International Publishing AG. 2017.

[13] Skansi S. Introduction to Deep Learning:From Logical Calculus to Artificial Intelligence, Springer International Publishing AG. 2018.

[14] Ajayi OO, Badrudeen AA, Oyedeji AI. Deep Learning Based Spectrum Sensing Technique for Smarter Cognitive Radio Networks. Journal of Inventive Engineering (JIET). 2021; 1(5), 64-77.

[15] Emara HM, El-Shafai W, Algarni AD, Soliman NF, El-Samie, FEA. A Hybrid Compressive Sensing and Classification Approach for Dynamic Storage Management of Vital Biomedical Signals. IEEE Access. 2023; 11, 108126-108151.

[16] Sanjaya WSM, Anggraeni D, Santika IP. Speech Recognition using Linear Predictive Coding (LPC) and Adaptive Neuro-Fuzzy (ANFIS) to Control 5 DoF Arm Robot. J. Phys.: Conf. Ser. 2018; 1090(012046), 1-10.

[17] Livieris IE, Pintelas E, Pintelas P. Gender Recognition by Voice Using an Improved Self-Labeled Algorithm. Mach. Learn. Knowl. Extr. 2019; 1, 492–503.

[18] Shaqra FA, Duwairi R, Al-ayyoub M, Shaqra FA, Duwairi R, Al-ayyoub M. Recognizing Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models. Procedia Computer Science. 2019; 151(2018), 37–44.

[19] Islam A, Sakib A. Bangla dataset and MMFCC in text-dependent speaker identification. Engineering and Applied Science Research. 2019; 46, 56–63.

[20] Shareef MS, Abd T, Mezaal YS. Gender voice classification with huge accuracy rate". TELKOMNIKA Telecommunication, Computing, Electronics and Control. 2020; 18(5), 2612-2617.

[21] Uddin MA, Pathan RK, Hossain MS, Biswas M. Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN. Journal of Information and Telecommunication. 2021; 1-16.

[22] Gevaert W, Tsenov G, Mladenov V, Member S. Neural Networks used for Speech Recognition. Journal of Automatic Control, University of Belgrade. 2010; 20, 1–7.