A Comparative Analysis of Machine Learning and Deep Learning Models for Gender Classification from Audio Speech

Mainul Islam[1*] and Md Nawab Yousuf Ali[1]

[1*]Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh.

*Corresponding author(s). E-mail(s): 2023-1-96-002@std.ewubd.edu;
Contributing authors: nawab@ewubd.edu;

Abstract

Recognizing gender from audio speech can improve human interactions with technologies. Though the human ear can identify the gender of a person from the sound of their voice, it can be quite complicated for an artificial intelligence (AI) system. Effective classification of gender from audio speech depends not only on the effective representation of the audio signal but also on the implementation of robust algorithms. In this research, we utilized Mel Frequency Cepstral Coefficients (MFCC) to represent the audio samples due to their effectiveness in capturing spectral characteristics, which highlight differences in the vocal tract structures between males and females. MFCC is considered one of the most efficient representations of audio signals that mimic human auditory systems. This study aims to analyze the effectiveness of various machine learning (ML) and deep learning (DL) methods in classifying gender from audio speech utilizing the MFCC feature representation. We experimented with several algorithms: SVM, KNN, stacking ensemble method, and LSTM. Three audio speech datasets were utilized to assess the performance of these algorithms. The best accuracies achieved in these datasets are 93.889%, 99.371%, and 94.558%. Furthermore, based on the findings of the experiment, this study proposes a framework for effective gender classification from audio speech.

Keywords: Gender classification, SVM, KNN, ensemble method, LSTM

1 Introduction

Speech, a medium of communication, is a fundamental way of expressing thoughts, emotions, and information. It allows us to exchange ideas and connect with others. The formation of speech can be attributed to the complex interplay of multiple physical organs, such as muscles in the mouth, throat, and lungs. These vocal organs generate sounds, which are then utilized to produce speech. The acoustic features of speech are impacted by multiple factors, such as person, gender, emotion, age, and numerous other factors. Based on the acoustic features of audio speech, it is possible to classify the gender of the voice. Features, such as pitch and formant frequencies, have been utilized for the gender classification from audio speech [1].

Over the last several years, there has been a substantial surge in the demand for spontaneous communication with technology. As technologies are incorporated at every step of our life, effective communication with technologies has become indispensable. Human-technology interactions through speech can be improved leveraging multiple methods, such as emotion recognition, gender classification, and speaker recognition.

Moreover, gender classification can be utilized to improve the user experience by customizing the interactions between humans and technology. Through precise identification of the speaker's gender, technology can modify its responses, language, and services to better accommodate personal preferences and requirements. Users will have more engaging and fulfilling interactions with interactive systems because of this personalization, which improves their overall effectiveness and usability.

ML and DL algorithms have proven to be powerful tools for analyzing audio speech, enabling the extraction of crucial information from speech signals. These approaches enhance the efficiency of speech analysis and facilitate the identification and interpretation of valuable insights embedded within audio data. These methods not only facilitate the classification of gender [2] but also enable the recognition of other valuable information such as emotions [3] and speakers [4].

In speech processing methods, DL is more effective in comparison to ML [5, 6]. Speech audio can be depicted as consecutive data, and some DL algorithms, such as Long Short-Term Memory (LSTM), are adept at handling such sequential data effectively. However, ML algorithms, such as K-Nearest Neighbors (KNN) and Support Vector Machines (SVM), have also produced promising results in gender classification from audio speech [7, 8].

Hızlısoy et al. [9] extracted Mel Frequency Cepstral Coefficients (MFCC) and mel spectrogram coefficients from the audio speech and compared the performance of several ML classification algorithms for gender recognition from audio speech. On the other hand, these studies [2, 10] proposed Bidirectional LSTM (BiLSTM)-based models for gender classification from audio speech. While Alamsyah and Suyanto [2] extracted MFCC to represent the audio, Alashban and Alotaibi [10] extracted multiple features, including MFCC, and utilized a combined feature representation for the audio speech.

Despite the numerous methods proposed for gender classification from audio speech, there remains a lack of comprehensive comparative analysis of these classification algorithms. Our study aims to present a comparison among the most effective methods for gender classification from audio speech. We utilized the MFCC feature to ensure that our comparison leverages the most established feature in the field, providing a solid basis for evaluating the performance of different algorithms. Moreover, MFCC can effectively capture spectral properties that reflect differences in vocal tract structure between males and females. Due to its effectiveness, MFCC is the most implemented feature representation in speech processing methods [11, 12].

We implemented both DL and ML classification algorithms for gender classification from audio speech, aiming to conduct a comparative analysis of their performance. As a deep learning algorithm, we utilized LSTM because it is designed to classify sequential data effectively. On the other hand, we separately applied two ML algorithms: SVM and KNN. SVM and KNN are the most commonly used machine learning algorithms for gender classification from audio speech. Furthermore, we applied a stacking ensemble method combining SVM and KNN. We conducted our experiment using three different audio speech datasets. Both ML and DL algorithms produced remarkable results in identifying the genders from the audio speech. However, the LSTM demonstrated superior performance compared to the ML algorithms. The contribution of this research is outlined below:

- Analyzing the performances of KNN, SVM, and LSTM in gender classification from audio speech.
- Implementing stacking ensemble method and assessing its performance.
- Comparative analysis of the ML and DL algorithms in gender classification from audio speech.

Based on the findings of our experiment where LSTM consistently produced the best accuracy, we propose a framework in Figure 1 for gender classification from audio speech. This framework can be integrated into any artificial intelligence (AI) based model to improve the user experience.

This paper is structured in the following manner: section 2 contains relevant literature, proposed method is described in section 3, datasets and experimental details are discussed in section 4 and section 5 respectively, section 6 contains experimental analysis and results. Finally, the paper is concluded in section 7.

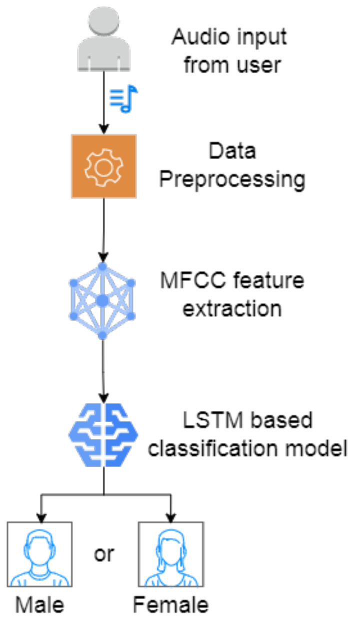

Fig. 1: Proposed framework for classifying gender based on audio speech.

2 Relevant literature

Various ML and DL classification algorithms, such as SVM [13], KNN [14], Random Forest (RF) [15], Convolutional Neural Network (CNN) [16], LSTM [17] and many others, have been implemented for gender recognition of audio speech. Each classification algorithm has its own strategy. On the other hand, several features, such as pitch, formants [1], MFCC [18], spectrogram [19], and linear prediction coefficients (LPC) [20], have been utilized to represent the audio speech. Uddin et al. [21] applied SVM and KNN for identifying gender from audio speech. The authors employed MFCC to represent the audio speech data. They utilized three datasets in their research, and across all three datasets, KNN outperformed SVM.

Hamdi et al. [22] proposed an ensemble classifier approach for gender identification from short speech audio samples, utilizing the Arabic Natural Audio Dataset (ANAD). They suggested a three-stage machine learning approach for feature optimization and achieved a high classification accuracy of 96.02% employing linear SVM classifier. In this paper [23], the authors implemented several machine learning algorithms, such as SVM, CatBoost, XGBoost, RF, Artificial Neural Networks (ANN), and many others, for gender classification. The best accuracy was obtained by CatBoost classifier, which obtained an accuracy of 96.4%.

Tursunov et al. [24] presented a model that combined CNN with multi-attention module. They employed the multi-attention module to capture essential features. Their model produced an accuracy of 96%. Mohammed and Al-Irhayim [25] applied Bidirectional Long-short Term Memory (BiLSTM) for gender recognition from audio speeches. They utilized MFCC to represent the audio samples. Their proposed model obtained an accuracy of 90.816%.

Samant et al. [26] developed a classification model for gender recognition based on voice features, evaluating the performance of several ML algorithms including SVM, KNN, RF, Decision Tree (DT), and Logistic Regression (LR). The study identified the best-performing model through analysis of accuracy, precision, recall, and F1-score, providing insights into effective methods for automatic gender identification. In this paper [27], the authors present a technique for gender identification in speech samples by leveraging the speech recognition process. They extracted 12 most relevant features from each audio sample to represent the audio. The study employs various machine and deep learning methods, such as RF, KNN, LR, DT, and CNNs, to classify these vectors into male and female categories. After evaluating the performance of these classifiers, the authors conclude that the CNN model is the most effective for gender classification.

Alashban and Alotaibi [10] proposed a model employing BiLSTM for gender identification. Their model incorporated several features representation of speech audio. Their suggested model achieved an accuracy of 91.76% in the Arabic-speakers model and 86.53% in the English-speakers model. In this paper [28], the authors implemented SVM and RNN for gender classification. They experimented with non-lexical speech features. They found that overlapping and lengthening are efficacious in gender recognition. They achieved an accuracy of 86.58% with SVM and 89.61% with RNN classifiers.

3 Proposed Method

Our proposed method is divided into the following parts: (1) MFCC feature extraction (2) implementing classification algorithms: SVM, KNN, LSTM and stacking ensemble method.

3.1 Feature extraction

We utilized Mel Frequency Cepstral Coefficients (MFCC) because of its effectiveness in capturing spectral characteristics that highlight differences in the vocal tract structures of males and females. The vocal tracts of men and women differ in length and shape, which influences formant frequencies and sound resonance. MFCC is adept at capturing these formant frequencies. Furthermore, MFCC emphasizes frequencies to which human ears are more attuned by using a mel-scale that emulates human auditory perception. Because of this, MFCC is especially effective at differentiating between the nuances of speech produced by men and women.

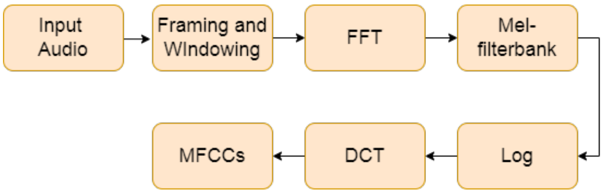

Fig. 2: Steps in calculating MFCC.

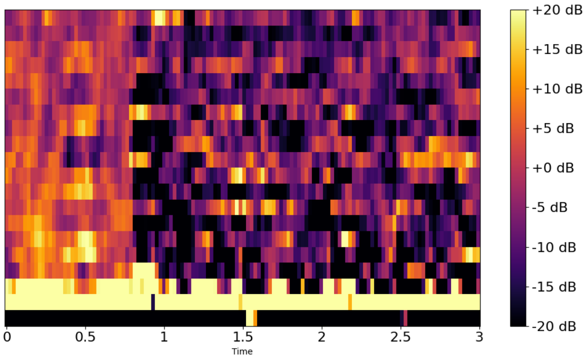

Fig. 3: MFCC visual representation of a female speech.

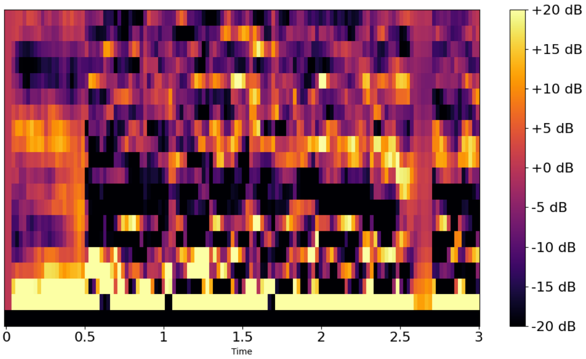

Fig. 4: MFCC visual representation of a male speech.

The steps in calculating MFCC are shown in Figure 2. After dividing the input audio into short frames, the Fast Fourier Transform (FFT) is applied to compute the power spectrum of each frame. Then the power spectrum is transformed by a mel-filterbank. Finally, the logarithm and Discrete Cosine Transform (DCT) are employed to produce the MFCC. Visual representations of MFCC of a female speech and a male speech uttering the same sentence are presented in Figure 3 and in Figure 4 respectively. However, in our experiment, we utilized a numeric representation of MFCC.

3.2 Support Vector Machine (SVM)

SVM operates by generating hyperplanes to separate different classes. The aim is to identify the most effective hyperplanes, which create the widest margin between the data of different classes. The margin refers to the separation between the hyperplane and the nearby data belonging to each class. For a two-class problem, the hyperplane is a line. On the other hand, the hyperplane is a plane for a problem that has more than two classes. In the case of not linearly separable data, various functions are used to transform the data so that it can be separated linearly. These functions are called kernels. Once the hyperplanes are evaluated, prediction can be made depending on which side of the hyperplane the new data falls.

3.3 K-Nearest Neighbors (KNN)

KNN determines the class of a given data by evaluating the most common class in its neighbors.

K is a hyperparameter that determines the number of neighbors to be considered. At first, the distances between the given data and the other data points are calculated. The distances are computed by employing various distance metrics, for example, Euclidean distance and Minkowski distance. Then the K closest neighbors is identified depending on the distance. Applying majority voting among the K closest neighbors, the class of the given data point is determined.

3.4 Long Short-Term Memory (LSTM)

LSTM is a deep learning model designed to work efficiently with the types of data where maintaining sequence is important. LSTM is capable of maintaining long-term dependencies effectively. It comprises multiple LSTM cells in a chain-like structure where the outcome from one LSTM cell serves as the input for the subsequent LSTM cell. The architecture of an LSTM cell is represented in Figure 5. The core concept behind this architecture is the horizontal line at top of the cell known as the cell state. By controlling the cell state, it is determined what portion of the previous information would be forgotten. LSTM works through the function of some gates: input gate, forget gate, and output gate.

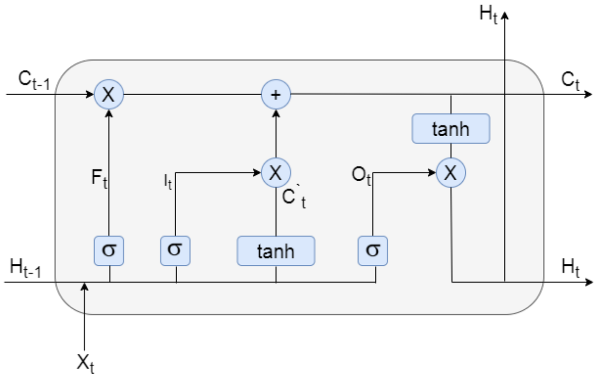

Fig. 5: Architecture of an LSTM cell.

How much of the prior information should be discarded is determined by the forget gate (F_t). On the other hand, the input gate determines how much new information would be included in the cell state. It is performed by computing two values: I_t and C'_t. The output gate (O_t) evaluates what portion of the cell state would be used to compute the output. It regulates how much information flows from the LSTM cell to the final output.

3.5 Stacking Ensemble Method

In the ensemble method, diverse algorithms are integrated to build a robust model. There are several types of ensemble methods, including stacking and bagging. We applied stacking ensemble method in our experiment. In stacking ensemble method, predictions produced by the base models are combined using a meta model. Initially, a diverse set of classification algorithms is selected as base models. The base models are selected in such a way that ensures the integration of diverse strategies. These base models are trained to generate individual predictions. Then these predictions are used as the inputs for the meta model, which produces the ultimate predictions. The objective is to merge the advantages of diverse algorithms to create a more effective model.

3.6 Proposed Architecture

We followed two architectures in our experiment. The architecture utilized for SVM, KNN, and LSTM is shown in Figure 6. The architecture used for the ensemble method is presented in Figure 7. However, the general steps are the same on both architectures which are as follows: audio input, feature extraction, feature standardization, train/test split, training the classification algorithms, and classifying the gender of the audio speech as male or female.

We represented each audio speech by 20 coefficients of MFCC. After standardization of the data, we divided it into train-test splits. We utilized 80% of the data for training and 20% of the data for testing. As ML classification algorithm, we separately applied SVM and KNN.

In the DL model employing LSTM, we applied an LSTM layer with 256 units and the ReLU activation function. After experimenting with several numbers of units in the LSTM layer, we selected 256 units based on our experiment. After this layer, we applied a dense layer with a sigmoid activation function.

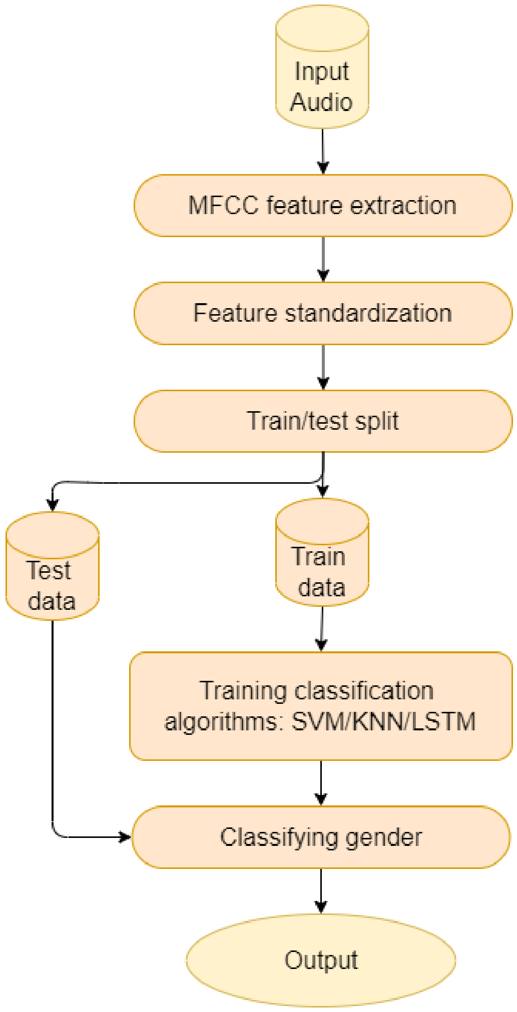

Fig. 6: Architecture for SVM/KNN/LSTM.

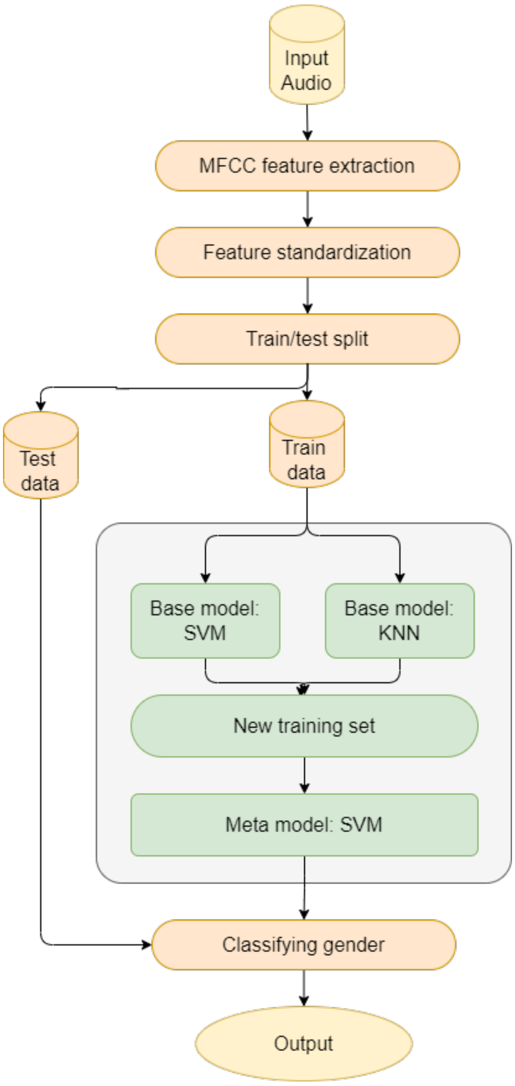

Fig. 7: Architecture for stacking ensemble method.

In the stacking ensemble method, we selected SVM and KNN as the base models. One of the strategies to improve the performance of the ensemble method is to combine algorithms that have different approaches of classification. SVM and KNN are the most utilized ML algorithms in gender recognition of audio speech. Each of these classification algorithms has its own strategies. SVM operates by creating hyperplanes to separate the data of different classes. on the contrary, KNN evaluates the K closest neighbors of a given data and applies majority voting among the selected neighbors to determine the class of the given data. As meta model, we employed SVM which was trained by the predictions produced by the base models. We utilized linear kernel in the SVM classifier. The meta model finally classified the gender of the test audio samples.

4 Datasets

We utilized three audio speech datasets in our experiments: KUET Bangla Emotional Speech (KBES) and Bangla Emotional Speech Recognition Dataset (BANSpEmo), and Bangla Speech Emotion Recognition (BanglaSER). These datasets were generated for speech emotion recognition. However, we applied them for gender recognition of audio speech. Bangla ranks among the most widely spoken languages globally. However, very few studies have been conducted utilizing Bangla audio datasets. Nevertheless, the proposed framework can be employed in any audio speech dataset for gender identification.

The KBES dataset contains 900 Bangla audio speeches produced by 35 speakers where 20 speakers were female, and 15 speakers were male. The speakers were of diverse age ranges. The dialogs of the audio speeches are almost unique and the length of each audio is 3 seconds. Of the audio samples, 450 were produced by male speakers and 450 were produced by female speakers. So, the target variable has a balanced distribution in this dataset. The details of this dataset have been discussed here [29].

The BANSpEmo consists of 792 Bangla audio samples created by 22 speakers where 11 speakers were female, and 11 speakers were male. The durations of the audio samples are of various lengths from 3 seconds to 12 seconds. Of the audio samples, 396 were produced by male speakers and 396 were produced by female speakers. The speakers uttered 12 sentences to produce the dataset. This dataset also has a balanced distribution of target variables. The details of this dataset have been discussed here [30].

The BanglaSER dataset contains 1467 audio samples generated by 34 speakers. Among the speakers, 17 were male and 17 were female. The duration of the audio samples ranges between 3 and 4 seconds. The dataset has a balanced distribution of male and female audio samples. In this paper [31], the details of the BanglaSER dataset have been described.

5 Experimental Details

We followed the same setup for the datasets. We captured 20 Mel-frequency cepstral coefficients (MFCC) from each audio speech. The MFCCs were extracted utilizing the Librosa Library in Python [32]. Of the data, 80% was used for training and 20% was used for testing. We employed a linear kernel in the SVM classifier. In KNN classifier, 5 nearest neighbors were used. In the stacking ensemble model, SVM and KNN had the same parameters.

In the LSTM layer, we applied 256 units and ReLu activation function. It was then followed by a dense layer with a sigmoid activation function. A batch size of 32 was utilized. We implemented the Adam optimizer and computed the model for 1000 epochs.

Across the datasets, we assigned the label ’1’ to the male class and ’0’ to the female class.

6 Experimental Analysis and Results

To assess the effectiveness of the algorithms, we utilized accuracy and precision score. The results are represented in Table 1, Table 2, and Table 3 for KBES, BANSpEmo, and EmoDB datasets respectively.

Table 1: Results obtained in the KBES dataset

Algorithms	Accuracy	Precision score
SVM	85.556%	85.979%
KNN	90.00%	90.00%
Ensemble method	92.222%	92.262%
LSTM	93.889%	91.566%

Table 2: Results obtained in the BANSpEmo dataset

Algorithms	Accuracy	Precision score
SVM	97.484%	97.510%
KNN	98.742%	98.742%
Ensemble method	99.371%	99.378%
LSTM	99.371%	98.837%

Table 3: Results obtained in the BanglaSER dataset

Algorithms	Accuracy	Precision score
SVM	85.714%	85.809%
KNN	91.156%	91.919%
Ensemble method	92.517%	92.546%
LSTM	94.558%	95.862%

In the KBES dataset, the best accuracy was achieved by LSTM, which obtained an accuracy of 93.889%. On the other hand, the best precision score was obtained by the ensemble

Table 4: Existing methods and their results for gender classification from audio speech

Reference	Description	Accuracy
Wani et al. [13]	SVM with time-frequency features. Two datasets were used: EmoDB and IITKGP-SEHSC	83.00% on IITKGP-SEHSC and 81.00% on EmoDB
Ali et al. [20]	Artificial neural network with MFCC	97.07%
Hamdi et al. [22]	SVM with feature optimization technique	96.02%
Tursunov et al. [24]	CNN with multi-attention module for feature extraction	96.00%
Mohammed and Al-Irhayim [25]	BiLSTM with MFCC feature	90.82%
Son et al. [28]	SVM and RNN with non-lexical speech features	86.58% using SVM and 89.61% using RNN
Chachadi and Nirmala [33]	Neural Network-based model with concatenated features	94.32%
Ghosh et al. [34]	Neural network-based model with combined features	90.74%
Badr and Abdul-Hassan [35]	CatBoost-based algorithm for features selection and SVM as classifier	89.62%

method. The ensemble method has produced better results compared to the individual ML algorithms (SVM and KNN) in this dataset. However, in the BANSpEmo dataset, the best accuracy was achieved by LSTM and ensemble method with an accuracy of 99.371%. Nevertheless, the best precision score was obtained by the ensemble method only. Also in this dataset, the ensemble method has obtained better results than the individual ML algorithms (SVM and KNN). In the BanglaSER dataset, LSTM again outperformed other algorithms with an accuracy of 94.558%. Unlike the other two datasets, the LSTM model achieved the best precision score on this dataset. Also, combining SVM and KNN in the ensemble method has produced an improved performance with an accuracy of 92.517%.

In general, we can observe consistent patterns of results across the three datasets where LSTM consistently achieved the highest accuracy. In the three datasets, the stacking ensemble method has produced better results in comparison to the individual ML algorithms (SVM and KNN). The ensemble method leverages the strategies of individual algorithms to improve the results. However, overall, the LSTM model consistently achieved the best accuracy across the three datasets. LSTM is designed to effectively classify discrete categories presented as sequential data. The chain-like structure of the LSTM model is tailored to capture the long-term dependencies in the sequential representation of data, such as the MFCC representation of audio data. LSTM models have produced remarkable results not only in gender recognition from audio speech but also in other speech-related techniques, for example, speech emotion recognition [36] and speaker recognition [37]. Existing methods and their results for gender classification from audio speech are presented in Table 4.

7 Conclusion

In conclusion, the goal of this research was to assess the effectiveness of ML and DL algorithms in gender classification from audio speech. Since our interactions with technologies are increasing every day, being able to extract information from audio speech is becoming increasingly essential. Gender classification from audio speech can improve the interactions between humans and technologies.

Both ML and DL have been utilized in previous studies for gender recognition from audio speech. However, our aim was to present a comparative analysis between ML and DL algorithms. Our study suggests that the ensemble method is more effective than the individual ML algorithms (SVM and KNN) in gender identification from audio speech. In BANSpEmo dataset, the ensemble method and LSTM obtained the best accuracy.

In the KBES and EmoDB datasets, however, the best accuracy was produced by the LSTM model. So LSTM model consistently produced the best accuracy across the three datasets. The best accuracies we obtained in BANSpEmo, KBES, and BanglaSER were 99.371%, 93.889%, and 94.558% respectively.

Future studies can apply the proposed methods to larger datasets. More ML and DL algorithms, such as Random Forest, BiLSTM, CNN, and others, can be implemented. Moreover, different ensemble approaches, such as bagging and boosting, can also be utilized.

Conflicts of Interest

The authors state that there are no conflicts of interest.

Data availability

The datasets utilized in this study are publicly available. The sources are given below:

KBES dataset: https://data.mendeley.com/datasets/vsn37ps3rx/4

BANSpEmo dataset: https://data.mendeley.com/datasets/rdwn4bs5ky/2

BanglaSER dataset: https://data.mendeley.com/datasets/t9h6p943xy/5

References

[1] P. Kumar, N. Jakhanwal, A. Bhowmick, and M. Chandra, "Gender classification using pitch and formants," in *Proceedings of the 2011 International Conference on Communication, Computing & Security*, pp. 319–324, 2011.

[2] R. D. Alamsyah and S. Suyanto, "Speech gender classification using bidirectional long short term memory," in *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pp. 646–649, IEEE, 2020.

[3] Y. B. Singh and S. Goel, "A systematic literature review of speech emotion recognition approaches," *Neurocomputing*, vol. 492, pp. 245–263, 2022.

[4] U. Ayvaz, H. Gürüler, F. Khan, N. Ahmed, T. Whangbo, and A. Bobomirzaevich, "Automatic speaker recognition using mel-frequency cepstral coefficients through machine learning," *CMC-computers materials & continua*, vol. 71, no. 3, 2022.

[5] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE access*, vol. 7, pp. 117327–117345, 2019.

[6] T. J. Sefara and T. B. Mokgonyane, "Emotional speaker recognition based on machine and deep learning," in *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pp. 1–8, IEEE, 2020.

[7] B. Jena, A. Mohanty, and S. K. Mohanty, "Gender recognition of speech signal using knn and svm," 2020.

[8] A. A. Abdulsatar, V. Davydov, V. Yushkova, A. Glinushkin, and V. Y. Rud, "Age and gender recognition from speech signals," in *Journal of Physics: Conference Series*, vol. 1410, p. 012073, IOP Publishing, 2019.

[9] S. Hızlısoy, E. Çolakoğlu, and R. S. Arslan, "Speech-to-gender recognition based on machine learning algorithms," *International Journal Of Applied Mathematics Electronics And Computers*, vol. 10, no. 4, pp. 84–92, 2022.

[10] A. A. Alashban and Y. A. Alotaibi, "Speaker gender classification in mono-language and cross-language using blstm network," in *2021 44th International conference on telecommunications and signal processing (TSP)*, pp. 66–71, IEEE, 2021.

[11] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE access*, vol. 9, pp. 47795–47814, 2021.

[12] M. La Mura and P. Lamberti, "Human-machine interaction personalization: a review on gender and emotion recognition through speech analysis," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pp. 319–323, IEEE, 2020.

[13] T. M. Wani, T. S. Gunawan, H. Mansor, S. A. A. Qadri, A. Sophian, E. Ambikairajah, and E. Ihsanto, "Multilanguage speech-based

gender classification using time-frequency features and svm classifier," in *Advances in Robotics, Automation and Data Analytics: Selected Papers from iCITES 2020*, pp. 1–10, Springer, 2021.

[14] I. Submitter, B. Jena, A. Mohanty, S. K. Mohanty, *et al.*, "Gender recognition and classification of speech signal," in *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*, 2021.

[15] K. Gupta, A. Sharma, and A. Mohapatra, "Comparative analysis of machine learning algorithms on gender classification using hindi speech data," in *Artificial Intelligence and Speech Technology*, pp. 363–370, CRC Press, 2021.

[16] K. Chachadi and S. Nirmala, "Gender recognition from speech signal using 1-d cnn," in *Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2021*, pp. 349–360, Springer, 2022.

[17] G. R. Nitisara, S. Suyanto, and K. N. Ramadhani, "Speech age-gender classification using long short-term memory," in *2020 3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 358–361, IEEE, 2020.

[18] K. D. A. Danuwar, K. Badal, S. Karki, S. Titaju, and S. Shrestha, "Nepali voice-based gender classification using mfcc and gmm," in *International Conference on Machine Intelligence and Signal Processing*, pp. 233–242, Springer, 2022.

[19] M. S. Jitendra and Y. Radhika, "Singer gender classification using feature-based and spectrograms with deep convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021.

[20] Y. M. Ali, E. Noorsal, N. F. Mokhtar, S. Z. M. Saad, M. H. Abdullah, and L. C. Chin, "Speech-based gender recognition using linear prediction and mel-frequency cepstral coefficients," *Indonesian J Electric Eng Comput Sci*, vol. 28, no. 2, pp. 753–761, 2022.

[21] M. A. Uddin, M. S. Hossain, R. K. Pathan, and M. Biswas, "Gender recognition from human voice using multi-layer architecture," in *2020 International conference on innovations in intelligent systems and applications (INISTA)*, pp. 1–7, IEEE, 2020.

[22] S. Hamdi, A. Moussaoui, M. Oussalah, and M. Saidi, "Gender identification from arabic speech using machine learning," in *International Symposium on Modelling and Implementation of Complex Systems*, pp. 149–162, Springer, 2020.

[23] S. R. Zaman, D. Sadekeen, M. A. Alfaz, and R. Shahriyar, "One source to detect them all: gender, age, and emotion detection from voice," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 338–343, IEEE, 2021.

[24] A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms," *Sensors*, vol. 21, no. 17, p. 5892, 2021.

[25] A. A. Mohammed and Y. F. Al-Irhayim, "Speaker age and gender estimation based on deep learning bidirectional long-short term memory (bilstm)," *Tikrit Journal of Pure Science*, vol. 26, no. 4, pp. 76–84, 2021.

[26] S. Samant, S. Prajapati, R. Chaudhary, A. Rathi, S. Arya, *et al.*, "Unveiling gender from speech: An investigation into acoustic features for accurate gender detection," in *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, vol. 6, pp. 1828–1833, IEEE, 2023.

[27] H. Q. Jaleel, J. J. Stephan, and S. A. Naji, "Gender identification from speech recognition using machine learning techniques and convolutional neural networks," *Webology*, vol. 19, no. 1, pp. 1666–1688, 2022.

[28] G. Son, S. Kwon, and N. Park, "Gender classification based on the non-lexical cues of emergency calls with recurrent neural networks (rnn)," *Symmetry*, vol. 11, no. 4, p. 525, 2019.

[29] M. M. Billah, M. L. Sarker, and M. Akhand, "Kbes: A dataset for realistic bangla speech emotion recognition with intensity level," *Data in Brief*, vol. 51, p. 109741, 2023.

[30] M. G. Hussain, M. Rahman, B. Sultana, and Y. Shiren, "Banspemo: A bangla emotional speech recognition dataset," *arXiv preprint arXiv:2312.14020*, 2023.

[31] R. K. Das, N. Islam, M. R. Ahmed, S. Islam, S. Shatabda, and A. M. Islam, "Banglaser:

A speech emotion recognition dataset for the bangla language," *Data in Brief*, vol. 42, p. 108091, 2022.

[32] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python.," in *SciPy*, pp. 18–24, 2015.

[33] K. Chachadi and S. Nirmala, "Voice-based gender recognition using neural network," in *Information and Communication Technology for Competitive Strategies (ICTCS 2020) ICT: Applications and Social Interfaces*, pp. 741–749, Springer, 2022.

[34] S. Ghosh, C. Saha, and N. Molakathaala, "Neuragen-a low-resource neural network based approach for gender classification," *arXiv preprint arXiv:2203.15253*, 2022.

[35] A. A. Badr and A. K. Abdul-Hassan, "Catboost machine learning based feature selection for age and gender recognition in short speech utterances.," *International Journal of Intelligent Engineering & Systems*, vol. 14, no. 3, 2021.

[36] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021.

[37] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.