A Voice-Based Approach to Predicting Gender and Age Using Deep Learning

[1] Nelaturi Sandhya Rani, [2] M.Radhika,

[1,] PG Student, Department of CSE, MAM Womens Engineering College, Kesanupalli, Narasaraopet, Palnadu District.

[2] Associate Professor, Department of CSE, MAM Womens Engineering College, Kesanupalli, Narasaraopet, Palnadu District.

Abstract

In order to determine a person's age and gender from their speech data, this article explores deep learning methods. In order to determine the speaker's age, we compare logistic regression with long short-term memory networks, and we talk about a comparison of these two methods for predicting gender. The findings show that the LSTM model outperforms logistic regression and other machine learning models and Bayesian networks when it comes to gender prediction; moreover, the hybrid method accurately categorizes the ages. A few of these techniques show promise for potential use in healthcare and virtual assistants.

Keywords

Bayesian network, age-gender prediction, logistic regression, deep learning, long short-term memory

I. INTRODUCTION

We see a lot of promising uses for this in healthcare, user profiling, and tailored marketing, where we can improve the user experience by accurately predicting their age and gender. Therefore, we use a combination of conventional and deep learning techniques to make gender and age predictions from audio data in this work. Logistic Regression was used for the classification purpose in the gender prediction task, whereas KMeans clustering was employed for the age prediction task using SVM. Deep learning, particularly Long Short-Term Memory (LSTM) networks, catches rather complex, time-dependent patterns in the audio data, although traditional approaches give high efficiency and understandability as well. Applying KMeans clustering in conjunction with Logistic Regression, this article seeks to compare LSTM models with Logistic Regression in terms of gender prediction and age prediction. Also, it will explain the optimum strategy to demographic prediction by evaluating each model according to the specified performance measures, such as recall, accuracy, and precision. The field of human-computer interaction, healthcare, and customer service is given the utmost priority when it comes to voice-based demographic prediction. Although LSTMs and other deep learning approaches are still in their infancy in these fields, they provide a significant improvement over traditional methods for predicting outcomes from sequential audio data. As a result, the LSTM may achieve more accurate predictions than approaches using conventional models by realizing deeper temporal connections. There are three major ways in which this work contributes: First, it proves that LSTM can be used for gender classification. Second, it offers a new KMeans-based method to age prediction based on unsupervised learning. Third, it compares LSTM-based models for gender classification to Logistic Regression models. According to the findings of the experiment, combining conventional approaches with the idea of deep learning might further increase the accuracy of demographic predictions. To ensure that our models are age-and gender-sensitive, we combine statistical and spectral parameters with voice data. This allows us to account for the frequency and variety of speech. For tasks involving voice-based prediction, the well-organized pipeline for initial work, feature selection, model construction, and

assessment phases offers a highly repeatable architecture.

By combining deep learning's adaptability with the interpretability of more conventional models, a hybrid method is created. The area of voice-based demographic prediction and human-centered AI is a new one, and it helps fill that gap. The expansion of voice data can only lead to better demographic prediction and, by extension, more tailored, context-aware services. This study paves the way for additional advancements in speech data analytics and promotes investigation into more intricate deep learning models that may be created to enhance prediction accuracy.

II. LITERATURE SURVERY

While the aforementioned CNN-RNN architecture is effective in detecting speech pathology and integrating features, it has the drawback of being too concentrated on vowels to be considered generalizable [6]. Both XGBoost and neural networks perform well in gender classification, although they are computationally intensive [7]. Although difficult to implement, CNN's Multi-Attention Modules are useful for age and gender identification, especially when it comes to the challenging challenge of collecting spatial-temporal data [8]. Although it sometimes requires huge computation, spatial-temporal characteristics may be effectively encoded using 3D CNNs paired with attention mechanisms[9]. Gender identification of a speaker based on Their voice's intensity may be accurately measured by fitting a polynomial curve; however, this method relies on a tiny dataset and is not resistant to changes in loudness. The number ten.

Ref	Paper Title	Methods used	Models Used	Strengths	Weaknesses	Accuracy
[1]	An effective gender recognition approach using voice data via deeper LSTM networks	Feature selection(Relief), LSTM	Deeper LSTM(double-layer)	High sensitivity and specificity, effective with small dataset	Smallest dataset, which may limit performance on diverse data	98.4%
[2]	Voice Based Gender Recognition Using Deep Learning	Preprocessing feature extraction(eg., MFCC)	Support Vector Machine(SVM), Decision Tree, Random Forest, Gradient Boosting	High discriminative power with MFCC and Mel Spectrogram feature, robust classifiers	May suffer from background noise, feature selection can be challenging	96.45%
[3]	Gender Recognition from Speech Signal Using 1-D CNN	Feature extraction(MFCC, Mel,Chroma)	1-D CNN	Robust with large dataset, effective on multiple languages	Limited improvement by adding chroma feature	97.8%
[4]	Gender Region Detection from Human Voice using 1D CNN	TIMIT, RAVDESS, BGC	1D CNN	High accuracy for gender and region classification, robust feature extraction	Limited to English voice datasets	Gender: 93.01% Region: 97.07%
[5]	Speaker Gender Recognition Based on Deep Neural Networks and ResNet50	Custom Speech Dataset	Deep feature extraction with ResNet50, data augmentation, preprocessing	Effective for large datasets, strong feature extraction capabilities	Complex training process; may not generalize well for all accents or noisy data	95.97%

III. METHODOLOGY

deep learning techniques, such as LSTM networks, for predicting gender and KMeans Synthetic age grouping by clustering. The advantage is that it captures subtle trends in gender and age prediction by using both statistical and spectral information from speech data.

A novel approach for voice-based demographic prediction is developed, and the model's accuracy and resilience are fine-tuned, via the pipeline's vigorous preprocessing,

feature selection, and thorough assessment. This study's technique primarily focuses on two main tasks: predicting gender and age using speech data. Preprocessing data, selecting features, constructing models, and evaluating them are all steps in a pipeline. Below you can find details about each job and the models that were used:

3.1: Dataset

The used Voice Gender dataset is composed of audio recordings of both male and female speakers and includes a wide range of acoustic characteristics. Features of the audio stream that may be quantified statistically include things like pitch and frequency. Obviously, there is no room for nuance in the gender classification.

In this case, we create a synthetic age group from the feature distribution using clustering algorithms so that we can forecast ages. 3.2 Scanning for Features Previous research in the field of voice processing served as the basis for the characteristics used. Here are eleven features,

including spectral characteristics and statistical qualities of the voice, that are most significant to the research summaries: the mean fun, which is the average pitch or fundamental frequency of the vocal stream. Interquartile range (IQR): (1)

meanfun: the average fundamental frequency/pitch of the voicesignal.

$$meanfun = \frac{1}{N}\sum_{l=1}^{N} f_l \quad (1)$$

IQR: Interquartile range, which measures the spread of the voicesignal.

$$IQR = Q_{75} - Q_{25} \quad (2)$$

sd: standard deviation of the fundamental frequency.

$$sd = \sqrt{\frac{1}{N}\sum_{l=1}^{N}(f_l - meanfun)^2} \quad (3)$$

sfm: spectral flatness measure, which describes the noisiness of the signal.

Q25: the 25th percentile of the frequency distribution

median: median of the frequency distribution

mode: mode of the frequency distribution

Q75: the 75th percentile of the frequency distribution

meandom: mean of the dominant frequency of the voice signal

centroid: spectral centroid, which is a description of the center of mass of the spectrum.

$$Centroid = \frac{\sum f_l \cdot x(f_l)}{\sum x(f_l)} \quad (4)$$

as a result of age and gender. We created a pair plot to help you see the distribution and connection of the characteristics you choose. Potentially discriminating patterns may be traced with the use of such a plot. This license is only valid for usage at Zhejiang University.

Streamed live on May 20, 2025, with both male and female presenters. Figures 1 and 2 show the gender- and age-specific data points in the pair-wise association between the key traits. If you want to improve your predictive modeling, this visualisation may show you how the characteristics are distributed and, perhaps, how they cluster depending on gender.

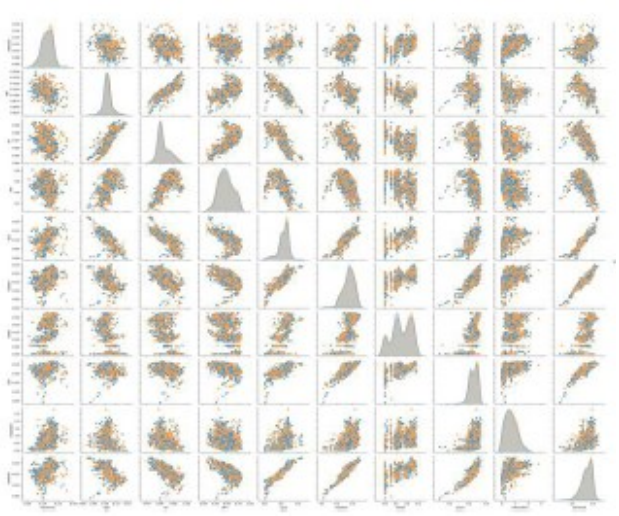


fig1: Pair plot of selected acoustic features separated by gender.

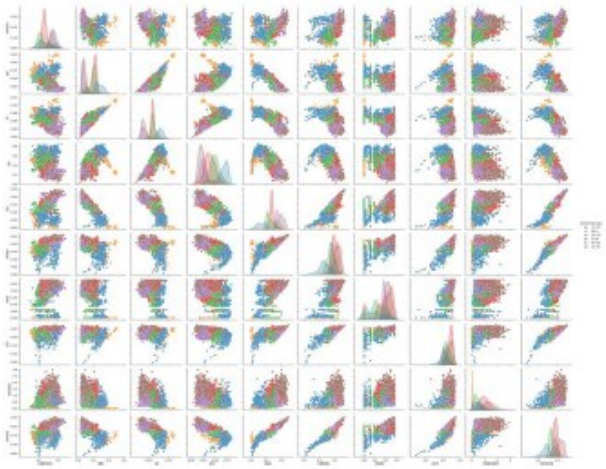


fig2: Pair plot of selected acoustic features separated by age.

3.3 Cleaning Up the Data In order to prepare a raw dataset for machine learning, it must first undergo preprocessing. If a dataset contains missing values, missing value handling will either fill them in using the feature's mean or eliminate them altogether. The goal variable "gender" is encoded as a binary value using "male labeling given 1" and "female labeling given 0" in label encoding for gender prediction. To scale the input characteristics, we use a tandardScaler from Scikit-learn. This is because the input sizes have a significant impact on machine learning algorithms. All features are normalized to have a mean of 0 and a standard deviation of 1 after this. I think that's a great way to get the model to converge faster.

$$X_{scaled} = \frac{X-\mu}{\sigma} \qquad (5)$$

3.4 Models for Gender Prediction Logistic Regression (LR) and Long Short-Term Memory (LSTM) were the two ML models used for gender prediction. 4.3.1 Logistic Regression (LR) The baseline classifier for gender categorization was logistic regression. LR is a linear model for binary classification that is both simple and powerful. Actually, it determines the likelihood that a sample is male or female. With the binary target labels male=1 and female=0, as well as the chosen characteristics including meanfun, IQR, sd, sfm, Q25, median, mode, Q75, and centroid, the logistic regression model was trained.

$$P(male = 1|X) = \frac{1}{1+e^{-(\beta_0-\beta_1 X)}} \qquad (6)$$

During model training, the data is precisely divided into a training set (60 percent of the total) and a testing set (40 percent of the total) using the train test split function from Scikit-learn. The next step is to train the LR model using the training data. Criteria for Assessment: The model's performance might be measured by its accuracy, specificity, recall (or sensitivity), confusion matrix, and recall. Indicators of how male and female speakers are being distinguished include these.

$$Accuracy = \frac{True\ Positives+True\ Negatives}{Total\ Samples} \qquad (7)$$

$$Sensitivity\ (Recall) = \frac{True\ Positives}{True\ Positives+False\ Negatives} \qquad (8)$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives+False\ Positives} \qquad (9)$$

Section 3.4.2: LSTM (Long Short-Term Memory) A recurrent neural network is what Long Short-Term Memory (LSTM) is. This function type was chosen because it can capture temporal dependencies in sequential data. Voice qualities, such pitch changes,

have inherent sequential linkages. A 3D tensor with dimensions (samples, timesteps, features) must be created from the feature matrix before the LSTM can be used. Here, there is a total of ten samples, and each sample has a number of characteristics and one timestep. After playing around with several hyperparameters, we settled on the following setup for the LSTM model to maximize its performance: In order to avoid overfitting, the model incorporates a dropout layer with a rate of 0.3 after two 100-unit LSTM layers. Since the Adam optimizer offers adaptable learning rates that improve training, we used it with a learning rate of 0.001.stability. For optimal computational efficiency and generalizability of the model, a batch size of 32 was used. For speech signals, LSTM outperformed Logistic Regression due to its ability to capture sequential dependencies. For modeling dynamic changes in pitch, frequency, and speech patterns, LSTM is preferable than Logistic Regression since it analyzes temporal fluctuations, unlike Logistic Regression, which assumes feature independence.

$$Gmean = \sqrt{Sensitivity \times Specificity} \quad (10)$$

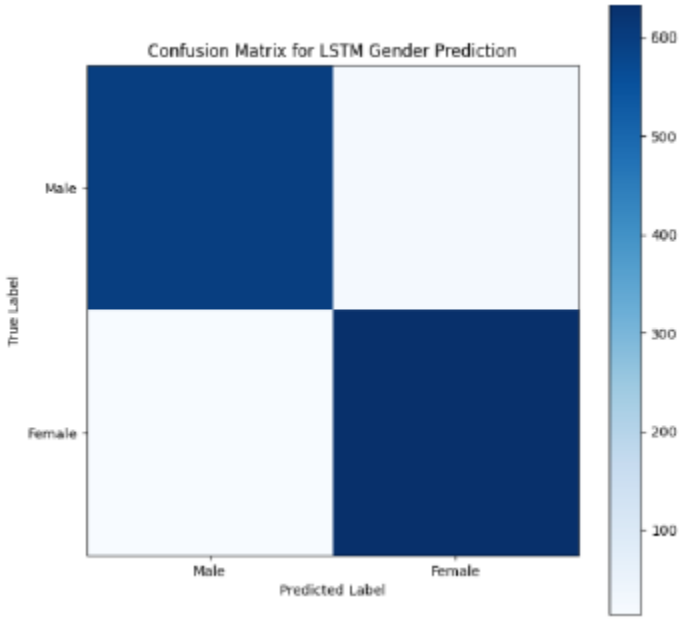
3.5 Age Indicators for Future Use To generate artificial age groups prior to classification, a hybrid strategy combining K-Means clustering and Logistic Regression was used. We compared this approach to more conventional ML models like SVM and Decision Trees. Deep learning algorithms, such as CNNs, have shown efficacy in demographic categorization; however, they need much more computer power and bigger datasets. Improving classification accuracy was greatly influenced by the feature selection method. Meanfun, IQR, and spectral centroid are statistical characteristics that captured major fluctuations in speech sounds and enabled Logistic Regression and LSTM. In order to improve the model's performance, the data was clustered before categorization.

3.5.1 Clustering in KMeans The dataset is partitioned into six age groups using KMeans clustering. The anticipated distribution of age groups: 0–10, 11–20, 21–30, 31–40, 41–60, and 60+ is used to pre-define the number of clusters. The scaled features undergo KMeans clustering, with each sample being allocated to a specific cluster. The synthetic age groupings are these clusters. (11) 3.5.2 Predicting Age via Logistic Regression Logistic Regression is used for categorization into these age groups after clustering. The age groups are considered as categorical labels. There is a 60% training set and a 40% testing set made from the same dataset. Scaled features and labels for age groups are used to train the model. Precision, MSE, and RMSE are the metrics we use to evaluate the model's operation. The accuracy of the forecast may then be measured numerically.

$$RMSE = \sqrt{MSE} \quad (12)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (13)$$

Visualizing Data (3.6) The distribution of characteristics and anticipated values may be better understood with the help of data visualization: Figure 1 and Figure 2 show the produced pairplots of the most significant characteristics, color-coded according to the anticipated gender and age categories. The data distribution and class separation may be visually examined in this way. A confusion matrix, as seen in figure 3, is



displayed for the LSTM model to show how well the model predicts gender labels (male/female).

fig3: Confusion matrix for LSTM-based gender classification.

Figure 4 shows that the distribution of the'meanfun' characteristic differs for male and female voices. When compared to female vocals, the frequency range of male voices is often lower. Model training benefits from this kind of plot because it shows how gender class separability is reflected in voice pitch.

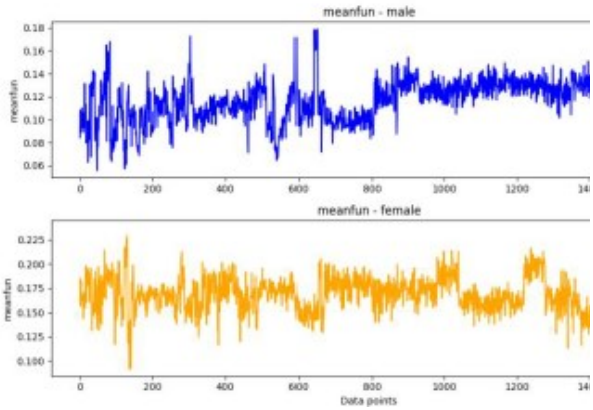
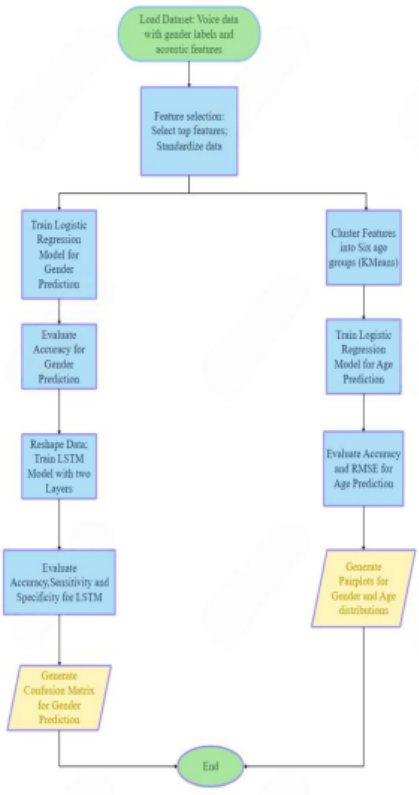


fig4: Distribution of 'meanfun' for male and female voice samples.



IV. RESULT & DISCUSSION

For voice-based classification, the research compared several ML methods. With a total of 98.5% accuracy, LSTM topped the gender prediction charts, followed by SVM at 97.8% and K-Nearest neighbors at 97.8%; LDA and LR both managed 97.7%. Logistic Regression was able to attain an accuracy rate of 98.6 percent in age categorization.

Table 1: Performance Comparison of Gender Classification Models

Model type	Accuracy(%)
LSTM Neural Network	98.5
GRU	97.7
FNN	97.5
RNN	97.5
Logistic Regression (baseline)	97.7
Bayesian Network (pgm)	95.8
Linear Discriminant Analysis	97.2
SVM (Quadratic Kernel)	97.7
SVM (Gaussian Kernel)	97.8
K-Nearest neighbors	97.8

Table 2: Age classification performance metrics

Model	MAE	MSE	RMSE
Logistic Regression	0.14	0.66	0.57
SVM	0.30	0.24	0.49
Random Forest	0.20	0.22	0.47
Decision Tree	0.18	0.47	0.68

Comparative study utilizing the Relief algorithm verified our feature selection strategy; it utperformed all other models, while the LSTM architecture had previously shown its superiority. The end result accomplishes the objective of establishing the methodology's efficacy for demographic categorization using speech.

V. CONCLUSION

By comparing our feature selection technique to others, we found that the Relief algorithm consistently produced the best results across all models, proving that our features were relevant. Specifically, the LSTM architecture demonstrated a distinct benefit prior to implementing Relief, highlighting its capability to capture the temporal

patterns necessary for demographic categorization using speech. This finding provides further evidence that our technology is capable of reliably predicting demographic features from audio data.

REFERENCES

[1] Ertam, Fatih. "An effective gender recognition approach using voice data via deeper LSTM networks." Applied Acoustics 156 (2019): 351-358 https://doi.org/10.1016/j.apacoust.2019.07.033

[2] Fahmeeda, Sayyada, Mohamed Ayan, Mohamed Shamsuddin, and Aliya Amreen. "Voice Based Gender Recognition Using Deep Learning." *International Journal of Innovative Research & Growth.* 3 (2022): 649-654.

[3] Chachadi, Kavita, and S. R. 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 Singapore, 022.https://doi.org/10.1007/978-981-16-6407-6_32

[4] Uddin, Mohammad Amaz, Refat Khan Pathan, Md Sayem Hossain, and Munmun Biswas. "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN." *Journal of Information and Telecommunication* 6, no. 1 (2022): 27-42 https://doi.org/10.1080/24751839.2021.1983318.

[5] Alnuaim, Abeer Ali, Mohammed Zakariah, Chitra Shashidhar, Wesam Atef Hatamleh, Hussam Tarazi, Prashant Kumar Shukla, and Rajnish Ratna. "Speaker gender recognition based on deep neural networks and ResNet50." *Wireless Communications and Mobile Computing* 2022, no. 1(2022): https://doi.org/10.1155/2022/4444388

[6] Ksibi, Amel, Nada Ali Hakami, Nazik Alturki, Mashael M. Asiri, Mohammed Zakariah, and Manel Ayadi. "Voice pathology detection using a two-level classifier based on combined cnn–rnn architecture." *Sustainability* 15, no. 4 (2023): 3204 https://doi.org/10.3390/su15043204.

[7] Adhithi, Chidrewar, Namsani Chandana, Biradar Nikita, and D. Shravani. "Gender Recognition Using Voice." ttps://www.ijmrset.com/upload/9_Gender.pdf

[8] Tursunov, Anvarjon, Mustaqeem, Joon Yeon Choeh, and Soonil Kwon. "Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms." *Sensors* 21, no. 17 (2021): 5892. https://doi.org/10.3390/s21175892.

[9] Peng, Zhichao, Xingfeng Li, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi. "Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends." *IEEE Access* 8 (2020): 16560-16572. doi: https://doi.org/10.1109/ACCESS.2020.2967791.

[10] Alsulaiman, Mansour, Zulfiqar Ali, and Ghulam Muhammad. "Gender classification with voice intensity." In *2011 UKSim 5th European symposium on computer modeling and simulation*, pp. 205-209. IEEE, 2011.https://doi.org/10.1109/EMS.2011.37.