Science World Journal Vol. 20(No 1) 2025 https://dx.doi.org/10.4314/swj.v20i1.18
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

A DEEP LEARNING MODEL FOR GENDER RECOGNITION USING VOICE DATA

[1]Aisha Kabir, *[2]Muhammad Aminu Ahmad, [2]Ahmad Abubakar Aliyu, [1]Saadatu Abdulkadir, [1]Abubakar Ahmed Muazu

[1]Department of Informatics, Kaduna State University, Kaduna, Nigeria
[2]Department of Secure Computing, Kaduna State University, Kaduna, Nigeria

*Corresponding Author Email Address: muhdaminu@kasu.edu.ng

ABSTRACT
Gender recognition using speech signals has become essential due to the advancement in digital technology and the need for computer systems to be able to classify gender using voice information. Numerous studies have been conducted with an emphasis on enhancing feature extraction and development of better classifiers for gender recognition based on speech. Out of all the different kinds of models developed, the LSTM model yields the greatest results. Additionally, for various signal to noise ratios, the LSTM model showed outstanding generalization performance. However, LSTM models use feed-forward neural networks that has limitations in capturing frequency and temporal correlations. This paves the way for further research into alternate recurrent-network techniques, which have been demonstrated to handle contextual information better, in order to achieve additional performance gains. The study improves gender recognition using a Bi-LSTM-LSTM architecture and voice data. The study adopts Relief-based method for feature selection. The results show that the BiLSTM-LSTM model achieved better gender recognition than LSTM-LSTM model at an accuracy of 99.30%, sensitivity of 99.60% and specificity of 99.00%. The BiLSTM model is successful in achieving higher accuracy and sensitivity values than LSTM at 1.00% and 2.20% respectively. The model also outperformed classical machine learning approaches (Fine Tree, K Nearest Neighbor, Linear Discriminant, Logistics regression and Support Vector Machine) in terms of accuracy at a minimum of 2.20% to a maximum of 05%. The comparative analysis of the classification performance shows that deep learning approaches are more successful in gender recognition than classical machine learning models.

Keywords: Gender recognition, Machine Learning, Deep Learning, BiLSTM, Speech Recognition

INTRODUCTION

Speech is one of the most common and important ways that people can express their feelings, thoughts, and intentions to one another. Gender identification can be accomplished with a speech and voice recognition system by selecting and integrating the appropriate features from voice data on a machine learning algorithm. Identifying a speaker's gender automatically offers a number of possible uses such as speech emotion recognition, human-machine interaction, telephone call sorting by gender classification, automated greetings, gender-specific sound muting, and audio/video categorization with tagging (Livieris et al., 2019). Furthermore, the advancement of voice recognition technology greatly increases people's comfort level while dealing with robots, computers and other devices (Nugroho, Noersasongko, & Santoso, 2019).

Machine learning is widely applied across various industries (Heidari, 2022). In an effort to create more accurate classifiers, recent research has concentrated on combining ensemble learning methods with the semi-supervised learning framework (Cances et al., 2022). In addition, deep neural networks offered tremendous gains in speech recognition. The speech recognition approach reported by Sharmin et al. (2020) to classify Bengali spoken digit and Pironkov, Wood and Dupont (2020) for automatic speech recognition using a convolutional neural network (CNN) outperformed the use of traditional machine learning approaches and ensemble techniques (Krishnakumar & Williamson, 2019). This shows the success of using deep learning in speech recognition.

Researcher also used deep learning models particularly Long Short-Term Memory (LSTM) and Convolutional Neural Network in voice detection. (Doukhan, et. Al., 2018) presented an open-source speaker gender detection framework for monitoring gender equality using CNN and evaluate it against GMM and I-Vector. The experimental results showed that the CNN method is sufficient to perform large-scale gender equality description. Krishnakumar and Williamson (2019) compared the performances of Boosted Deep Neural Networks, LSTM, CNN and Ensemble for voice activity detection. The results showed that LSTM has the best performance. In addition, Ertam (2019) presented a gender recognition technique using voice data via deeper LSTM networks. The technique was compared against SVM, KNN and LR. The results showed that LSTM has the best performance at 98.4% accuracy. These show the strength of LSTM models over CNN and classical machine learning techniques in voice activity detection. This is due to the generalization capability of LSTM models with both seen and unseen results at different signal to noise levels (Ye et al., 2022).

Furthermore, LSTM models use feed-forward neural networks that has limitations in capturing frequency and temporal correlations (Krishnakumar & Williamson, 2019). Thus, it is essential to explore the use of other technique that will enhance classification performance, such as BiLSTM hybridization because it handles contextual information better (Wubet and Lian, 2022). Therefore, this study presents an approach that uses Bi-LSTM hybridization to improve gender recognition using voice data because the backward propagation capability of BiLSTM preserves information from the future using hidden states (Greff et. Al., 2016).

MATERIALS AND METHOD

The approach developed for gender recognition using voice data is illustrated in Figure 1.

Science World Journal Vol. 20(No 1) 2025 https://dx.doi.org/10.4314/swj.v20i1.18
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

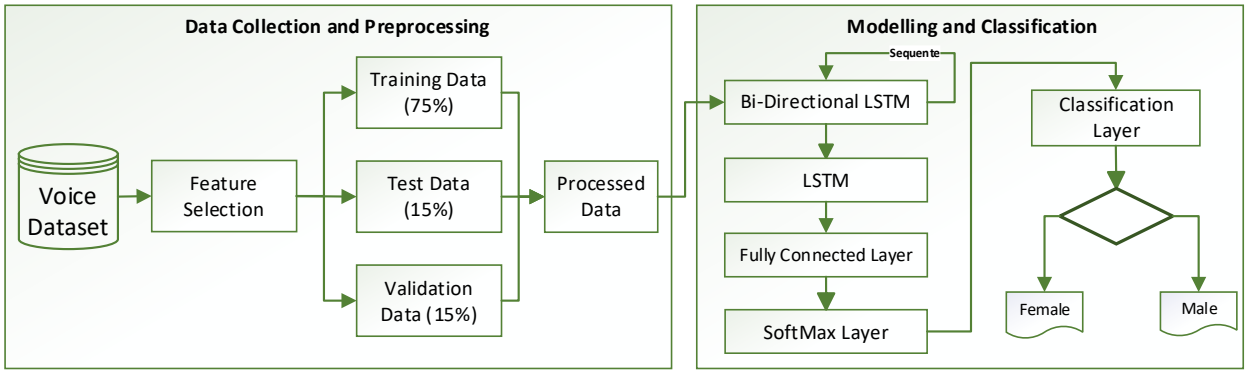

Figure 1: Gender classification using Bi-LSTM-LSTM deep architecture

The developed technique comprises two main components; Data Collection and Pre-processing and Modelling and classification. Initially, a collected voice data is pre-processed by cleaning, normalizing and selecting relevant features from the data. The data is then divided into three for training, testing and validation. The processed data is then passed to the modelling and classification component for gender identification and recognition.

Data Collection and Pre-processing

A public voice dataset (Becker, 2016) was collected, which comprises 3168 voice data; 50% male voice and 50% female voice. The dataset has 20 features as shown in Table 1. Relief-based feature selection method (Urbanowich, 2018) was used to get the best features based on their weight values as shown in Table 1. The most effective features are those with largest weight values based on the weight ranking. The top ten attributes are *meanfun, IQR, sd, sfm, Q25, median, mode, Q75, meandom* and *centroid*. After selecting the ten best features, the data was split into 70% for training, 15% for testing and 15% for validation. The processed data was used for classification task.

Table 1: Voice Dataset Attributes

ID	Feature	Description	Weight	Weight Ranking
1	meanfreq.	mean freq	-0.000292171	20
2	sd	standard deviation of freq	0.025104974	3
3	median	median freq	0.018300957	6
4	Q25	Quantile (1.)	0.020403385	5
5	Q75	Quantile (3.)	0.016572387	8
6	IQR	interquantile range	0.032888221	2
7	Skew	skewness	0.007771738	11
8	kurt	kurtosis	0.00335422	16
9	sp.ent	spectral entropy	0.003275844	17
10	sfm	spectral flatness	0.022487556	4
11	mode	mode freq.	0.018044391	7
12	centroid	frequency centroid	0.009659533	10
13	meanfun	fundamental freq. (average)	0.092036098	1
14	minfun	fundamental freq. (minimum)	0.001123091	18
15	maxfun	fundamental freq. (maximum)	0.000847013	19
16	meandom	Dominant freq. (average)	0.011101854	9
17	mindom	Dominant freq. (minimum)	0.006186568	13
18	maxdom	Dominant freq. (maximum)	0.004829646	14
19	dfrange	range of dominant freq.	0.004756876	15
20	modindx	modulation index	0.006769619	12
21	label	Class, Female (0) or male (1)		

Modelling and Classification

A family of artificial neural networks known as a recurrent neural network (RNN) uses feedforward neural networks, in which node connections create a directed graph along a temporal sequence. Variable length input sequences can be processed by RNNs using their internal state, or memory (Dupond, 2019). They can therefore be used for tasks like speech recognition.

A particular kind of RNN that is capable of learning long-term dependencies is the Long Short-Term Memory network (LSTM). It is an excellent method for dealing with the vanishing gradient issue and effective in capturing both long-term dependencies and non-linear relationships in complicated datasets. In lower layers of a deep network, LSTMs aid in preserving the error that can be back propagated over time (Bahad et al., 2019). A Bidirectional LSTM uses two LSTMs to an input sequence: one from the future to the past and one from the past to the future, which enhance model performance on sequence classification issues. Unlike LSTM models that are unidirectional, the backwards propagation capability of BiLSTM preserves information from the future using hidden states (Greff et. Al., 2016).

Science World Journal Vol. 20(No 1) 2025 https://dx.doi.org/10.4314/swj.v20i1.18
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Thus, a deep learning model was developed by combining Bi-LSTM and LSTM architectures. The bidirectional LSTM layer looks at sequence of features in both forward and backward directions, while LSTM is used to keep the information in memory, thereby forming a BiLSTM-LSTM classification model. Table 2 shows the parameters used to set up the model testing.

Table 2: Parameters Settings for the BiLSTM Model

Parameter	Settings
Input Layer	1
Bi-LSTM Layer	50
Fully Connected Layer	2
SoftMax Layer	1
Classification Layer	1
Max Epochs	4
Mini Batch Size	256
Verbose	False
Learn Rate Schedule	Piecewise

The number of hidden neurons for the BiLSTM layer was set to 50. The maximum epoch was set to four with a mini batch size of 256 to enable the network make four passes during training and 128 signals at a time. This will balance training speed and model convergence The verbose parameter was set to *"false"* and the learn rate schedule was set to "*piecewise*" to decrease the learning rate by a specified factor (0.1) every time a certain number of epochs (1) has passed. The BiLSTM model was implemented in MATLAB simulation environment.

Evaluation Metrics

This study uses a confusion matrix; True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) (Heydarian et. al. 2022) to evaluate the detection performance of the developed BiLSTM-LSTM model. True positive is when the data point's actual class is 1 (true) and its predicted class is also 1 (true). True negative is when the data point's actual class is 0 (false) and its predicted class is likewise 0 (false). False positive is when a data point's predicted class is 1 (true) but its actual class is 0 (false). False negative is when a data point's predicted class is 0 (false) and its actual class is 1 (true). The overall gender classification performance of the model was tested using accuracy, sensitivity and specificity as defined in Equations 1 to 3 (Naidu & Sibanda, 2023).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{Specificity} = \frac{TN}{FP + TN} \quad (3)$$

RESULTS AND DISCUSSION

This section presents the result obtained after evaluation. Figure 2 shows the results of the evaluation highlighting the parameters used for the BiLSTM-LSTM model.

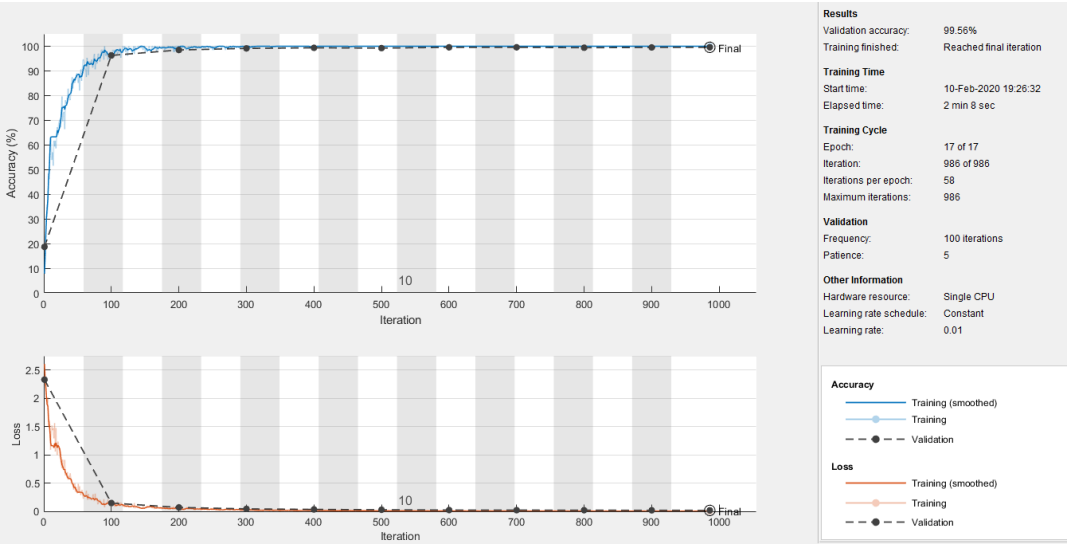

Figure 2: Training performance for the Bi-LSTM Model

The classification accuracy on each mini-batch is shown in the top subplot of the training-progress plot. This value usually rises near 100% as training proceeds effectively. The training loss, or the cross-entropy loss for each mini-batch, is shown in the bottom subplot. This value usually drops toward 0 as training proceeds well. The plots may fluctuate between values without showing an upward or downward trend if the training is not convergent. This oscillation indicates that neither the training loss nor the training accuracy is increasing. This condition can occur at the outset of training or after some preliminary progress in training accuracy. Changing the training options can often aid in the network's convergence. The network can learn better by decreasing the size of mini batch or initial learning rate though it may increase the training time.

The gender classification performance of the developed BiLSTM-LSTM model is presented in Table 3 using accuracy, sensitivity and specificity.

Table 3: Gender Classification Performance

S/N	Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Fine Tree	96.20	94.60	97.90
2	Linear Discriminant	96.60	97.40	95.70
3	Logistic Regression	96.50	96.00	96.90
4	Support Vector Machine	97.30	96.90	97.70
5	K Nearest Neighbour	97.60	97.40	97.70
6	LSTM-LSTM	98.40	97.40	99.50
7	BiLSTM-LSTM	99.30	99.60	99.00

The metrics were derived using confusion matrix results of the BiLSTM model (TP = 1568, TN = 1577, FP = 16, FN = 7) as specified in Equations 1 to 3. The table also shows the gender classification performances of some classical machine learning models (using Fine Tree, Linear Discriminant, Logistic Regression, Support Vector Machine and K Nearest Neighbor algorithms) and an LSTM model that used the same dataset.

Accuracy is the proportion of correct prediction made by a classification model. From the results, the Bi-LSTM-LSTM model achieved an accuracy of 99.30% while the LSTM-LSTM model has an accuracy of 98.40%. The two deep learning models have the highest gender classification performances in comparison with the classical machine learning models (with SVM and KNN having accuracy values of 97.30% and 97.60%, while Fine Tree, Linear Discriminant and Logistics regression have accuracy values of 96.20%, 96.60% and 96.50% respectively). This emphasizes the performance of deep learning approaches over classical machine learning approaches as reported by Ertam (2019).

Furthermore, the sensitivity (also referred to as recall or true positive rate) of a classification test shows how well a model can classify samples. The results show that the BiLSTM-LSTM model has the highest sensitivity value of 99.60%, which is followed by LSTM, KNN and Linear Discriminant models with 97.40% each, then SVM and KNN with 96.60% and 96.00% respectively. The Fine tree model has the lowest accuracy and sensitivity values of 96.20% and 94.60% respectively.

Finally, specificity measures of how well a classification test can identify true negatives, that is, the likelihood that the system detects female when the input is female speaker. The BiLSTM-LSTM mode achieved specificity of 99.00% which is slightly lower that the LSTM-LSTM model with a specificity of 99.50%. This is due to higher number of false positives for the BiLSTM-LSTM model. Fine Tree achieved 97.90%, Linear Discriminant achieved 95.70%, Logistic Regression achieved 96.90%, SVM and KNN achieved 97.70% each.

To conclude, the BiLSTM-LSTM model outperformed the LSTM-LSTM model in terms of accuracy and sensitivity by 0.90% and 2.20% respectively, but slightly performed lower than the LSTM-LSTM model by 0.50% as shown in Table 4. The developed model generally outperformed all the classical machine learning model in terms of accuracy with 3.10%, 2.70%, 2.80%, 2.00% and 1.70% higher values than that of Fine Tree, Linear Discriminant, Logistic Regression, SVM and KNN models respectively as shown in Table 4.

Table 4: BiLSTM-LSTM Model Performance Difference Compared to Other Models

S/N	Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Fine Tree	3.10	5.00	1.10
2	Linear Discriminant	2.70	2.20	3.30
3	Logistic Regression	2.80	3.60	2.10
4	SVM	2.00	2.70	1.30
5	KNN	1.70	2.20	1.30
6	LSTM-LSTM	0.90	2.20	-0.50

*Positive value shows the degree of how the BiLSTM-LSTM model outperformed a model and vice versa

The model also outperformed Fine Tree, Linear Discriminant, Logistic Regression, SVM and KNN models in terms of sensitivity by 5.00%, 2.20%, 3.60%, 2.70%, 2.20% respectively, and by 1.10%, 3.30%, 2.10%, 1.3% and 1.3% in terms of specificity respectively. Hence, this shows the success of Bi-LSTM-LSTM model in gender recognition using voice data over LSTM and other machine learning models.

Conclusion and Future Work

Classifying the age and gender of speakers is one of the challenging speech processing tasks because the accuracies obtained need improvement. An LSTM network performs exceptionally well in gender categorization, as demonstrated in the literature, showing that contextual information retention is the most appropriate approach for this problem. Additionally, for various signal to noise ratios, the LSTM model showed outstanding generalization performance on both seen and unseen data. This paves the way for additional research into different recurrent-network strategies. This study used a BiLSTM-LSTM network to classify gender using voice data. An LSTM network can learn long-term dependencies between time steps of a sequence. while a bidirectional LSTM layer can look at the time sequence in both forward and backward directions. The classification uses 10 features out of the 20-feature set in the voice based on the outcome of Relief-Based feature selection method. The BiLSTM-LSTM model classification achieved 99.30% accuracy, with sensitivity and specificity values of 99.60% and 99.00%, respectively. The developed model outperformed classical machine learning classifiers (Fine Tree, Linear Discriminant, Logistic Regression, SVM, KNN) in terms of accuracy, sensitivity and specificity. It also performed better than the LSTM model in terms of classification accuracy and sensitivity. The experimental results show that the deep learning approach gives better performance compared to classical machine learning approaches and Bi-LSTM has the potential to enhance gender classification using voice data.

The study used an existing voice data and focused on classification performance of deep learning models without considering computational complexity. Future work can assess the computational complexity of deep learning models on gender classification. This is to enable deployment on resource constraints devices and systems. In addition, multiple and diverse voice datasets can be used to determine the performance of classification models.

REFERENCES

Bahad, P., Saxena, P., & Kamal, R. (2019). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Computer Science, 165*, 74-82.

Cances, L., Labbé, E., & Pellegrini, T. (2022). Comparison of semi-supervised deep learning algorithms for audio classification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2022*(1), 23.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., & Meignier, S. (2018). *An open-source speaker gender detection framework for monitoring gender equality.* Paper presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annual Reviews in Control, 14*, 200-230.

Ertam, F. (2019). An effective gender recognition approach using voice data via deeper LSTM networks. *Applied Acoustics, 156*, 351-358.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems, 28*(10), 2222-2232.

Heidari, A., Navimipour, N. J., & Unal, M. (2022). Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review. *Sustainable Cities and Society*, *85*, 104089.

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *Ieee Access*, *10*, 19083-19095.

Krishnakumar, H., & Williamson, D. S. (2019). *A Comparison of Boosted Deep Neural Networks for Voice Activity Detection.* Paper presented at the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP).

Livieris, I. E., Pintelas, E., & Pintelas, P. (2019). Gender Recognition by Voice Using an Improved Self-Labeled Algorithm. *Machine Learning and Knowledge Extraction, 1*(1), 492-503.

Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference* (pp. 15-25). Cham: Springer International Publishing.

Nugroho, K., Noersasongko, E., & Santoso, H. A. (2019). *Javanese Gender Speech Recognition Using Deep Learning And Singular Value Decomposition.* Paper presented at the 2019 International Seminar on Application for Technology of Information and Communication (iSemantic).

Pahwa, A., & Aggarwal, G. (2016). Speech feature extraction for gender recognition. *International Journal of Image, Graphics and Signal Processing, 8*(9), 17.

Pironkov, G., Wood, S. U., & Dupont, S. (2020). Hybrid-Task Learning for Robust Automatic Speech Recognition. *Computer Speech & Language*, 101103.

Sharmin, R., Rahut, S. K., & Huq, M. R. (2020). Bengali Spoken Digit Classification: A Deep Learning Approach Using Convolutional Neural Network. *Procedia Computer Science, 171*, 1381-1388.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, 189-203.

Wubet, Y. A., & Lian, K. Y. (2022). Voice conversion based augmentation and a hybrid CNN-LSTM model for improving speaker-independent keyword recognition on limited datasets. *IEEE Access*, *10*, 89170-89180.

Ye, W., Jiang, Z., Li, Q., Liu, Y., & Mou, Z. (2022). A hybrid model for pathological voice recognition of post-stroke dysarthria by using 1DCNN and double-LSTM networks. *Applied Acoustics*, *197*, 108934.