International Journal of Advanced Trends in Multidisciplinary Research
ISSN: 3139-0986, Volume 2 Issue 7
July 2026, PP 1-7

Hybrid Speech-Based Gender Recognition and Age Classification Using LSTM Networks

S.Akhila[1], Ch.sathvika[2]
[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana, India.
[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana, India.

Abstract **– Due to its extensive usage in intelligent virtual assistants, healthcare systems, customer analytics, and human-computer interface, voice-based demographic prediction has become a significant field of study in speech processing. Developing personalised and context-aware services is made possible by accurately recognising demographic factors like age and gender from voice signals. For the purpose of determining an individual's age and gender from audio recordings, this research introduces a framework that combines deep learning with machine learning. To assess the benefits of sequential deep learning over traditional machine learning methods, we compare Logistic Regression's performance with that of an LSTM network for gender categorisation. Logistic Regression is used for multi-class classification after K-Means clustering is used to create synthetic age groups for age prediction. To accurately describe speaker-specific speech patterns, the suggested framework employs a number of spectral and acoustic parameters, such as statistical frequency descriptors, dominant frequency, interquartile range, spectral centroid, and mean fundamental frequency. In order to make the model more resilient and the predictions more accurate, we use thorough preprocessing, feature scaling, and feature selection. In terms of gender classification, the experimental evaluation shows that the LSTM model outperforms the baseline Logistic Regression model with a 98.5% accuracy rate, while the hybrid age prediction method successfully categorises synthetic age groups with high predictive performance. These results show that for intelligent applications in the real world, combining deep learning with traditional machine learning gives a scalable and dependable solution for voice-based demographic prediction.**

Keywords **– Speech-Based Gender Recognition, Age Classification, Long Short-Term Memory (LSTM), Deep Learning, Speech Signal Processing, Voice Biometrics, Audio Feature Extraction**

I. INTRODUCTION

The fast development of AI and speech processing technology has made the automatic detection of demographic features from human speech an important area of study. Physiological and behavioural traits of speakers, such as their gender, age, emotional state, and health issues, are reflected in the important acoustic information inherent in voice signals. Intelligent systems that can provide personalised services in many sectors, such as healthcare, biometric identification, customer assistance, virtual assistants, and adaptive human-computer interaction, may be created by extracting these features. Therefore, as a non-invasive and effective substitute for traditional identification techniques, voice-based demographic prediction has attracted a lot of interest.

For speech-based classification problems, traditional ML algorithms have been extensively used because to their computational efficiency, simplicity, and ease of implementation. When it comes to predicting gender, Logistic Regression is still among the most used binary classification methods. Traditional machine learning approaches, on the other hand, tend to treat input data as if they are completely independent, which means they could miss some of the subtle temporal correlations in voice signals. A model that can learn long-term temporal correlations is important because human speech shows sequential fluctuations in pitch, frequency, energy, and spectral properties across time.

The ability for deep learning to automate feature learning from sequential data has revolutionised voice analysis. When it comes to modelling temporal dependencies, Long Short-Term Memory (LSTM) networks stand out. Their memory cells and gating mechanisms keep contextual information intact throughout time sequences. Because of their superior ability to grasp dynamic speech patterns, LSTM networks are well-suited to gender categorisation tasks that include small temporal fluctuations influencing prediction accuracy, in contrast to classical classifiers. Also, when explicit age labels aren't given, age prediction may still be done using unsupervised learning methods like K-Means clustering, which effectively arrange speech samples into meaningful synthetic age groups before categorisation.

A hybrid framework for demographic prediction is introduced in this study for the analysis of speech recordings. It integrates Logistic Regression, Long Short-Term Memory (LSTM) networks, and K-Means clustering. To determine gender, we compare the results of Logistic Regression and Long Short-Term Memory (LSTM) models, and to determine age, we combine Logistic Regression with K-Means clustering. For accurate representation of speaker characteristics, the framework employs a number of carefully chosen acoustic features, such as spectral flatness measure (sfm), interquartile range (IQR), dominant frequency, standard deviation (sd), quartile frequencies, and spectral centroid. The computational efficiency and reliability of predictions are

further enhanced by using standard preprocessing methods, scaling features, and systematically evaluating models.

The main goal of this research is to find out how well voice-based demographic prediction works when traditional machine learning and deep learning are combined. The study sheds light on how to choose appropriate algorithms for speech-based age and gender estimation by comparing several prediction models under the same experimental circumstances. The experimental results show that future intelligent speech analysis systems may benefit from merging LSTM with classic machine learning methods, which improves the accuracy of demographic predictions.

II. LITERATURE SURVEY

Applications in intelligent speech interfaces, healthcare monitoring, biometric identification, and personalised human-computer interaction have garnered substantial interest to voice-based demographic prediction. In an effort to better determine a person's age and gender from audio data, researchers have looked into deep learning frameworks in addition to more traditional machine learning techniques. New research shows that deep neural networks can learn more complicated audio representations than classic classifiers, although traditional classifiers still have computational efficiency and understandable models. For voice-recorded automated gender recognition, Ertam suggested a more advanced Long Short-Term Memory (LSTM) architecture. The research demonstrated great sensitivity and specificity and obtained an exceptional classification accuracy of 98.4% by combining feature selection with the Relief algorithm. Despite the impressive performance of the suggested model, its performance may not be applicable to different speech settings due to the evaluation's use of a tiny dataset. Still, LSTM was shown to be a powerful deep learning model for sequential speech categorisation in the study.

In a separate but equally important work, Fahmeeda et al. explored voice-based gender detection using deep learning in conjunction with more conventional ML techniques including Decision Trees, Random Forests, Gradient Boosting, and Support Vector Machines (SVM). The significance of Mel-Frequency Cepstral Coefficients (MFCC) and feature extraction based on Mel spectrograms for enhancing gender categorisation accuracy was highlighted in their study. The scientists noted that while these discriminative acoustic characteristics greatly improved predictive accuracy, the model's resilience was still negatively impacted by background noise and the complexity of feature selection.

Another area that has shown success with deep convolutional architectures is speech-based demographic prediction. Using MFCC, Mel spectrogram, and Chroma characteristics, Chachadi and Nirmala presented a 1D-CNN for gender identification in multilingual voice datasets. They confirmed that CNN models could automatically extract discriminative auditory representations, and their trials showed that they could classify quite well. For the purpose of classifying gender and regional accents at the same time, Uddin et al. used a 1D-CNN architecture with the TIMIT, RAVDESS, and BGC speech datasets. Their results showed that convolutional neural network (CNN) feature extraction successfully caught gender-specific and regional speech features, even though the assessment was mostly conducted on English audio datasets.

Alnuaim et al. put forth a ResNet50–based deep neural network architecture for speaker gender recognition not long ago. In order to enhance classification performance on big voice datasets, their method included deep feature extraction, data enrichment, and thorough preprocessing. The research proved that deep residual learning can accurately represent speaker features; nevertheless, training deep architectures is computationally expensive, which makes it difficult to implement on systems with limited resources.

Few studies have focused on combining Logistic Regression, LSTM, and K-Means clustering into a single framework that can predict both gender and age at the same time, even though prior work has shown that deep residual networks, LSTM, and CNN all work well individually. To far, most approaches have either relied on gender categorisation or necessitated completely labelled age datasets. This work overcomes these shortcomings by creating synthetic age groups using K-Means clustering, using LSTM for sequential speech modelling, and using Logistic Regression for baseline gender classification. Then, it uses Logistic Regression for age classification. To improve the accuracy of demographic predictions while keeping computational efficiency high, this hybrid framework combines the best features of deep learning with those of conventional machine learning.

III. METHODOLGY

For the purpose of demographic prediction using speech recordings, the suggested methodology presents a hybrid prediction framework that integrates deep learning, conventional machine learning, and unsupervised learning approaches. Gender categorisation and age prediction are two separate but related functions of the framework. To forecast gender, we compare Logistic Regression with Long Short-Term Memory (LSTM) networks; to forecast age, we use K-Means clustering to create artificial age groups, and then we classify them using Logistic Regression. Preparing the dataset, designing features, preprocessing, training the model, making predictions, and evaluating performance make up the processing pipeline as a whole.

For these tests, researchers used the Voice Gender Dataset, a collection of audio recordings including male and female speakers. Various statistical and acoustic voice descriptors that capture the temporal and spectral aspects of speech signals are included in the collection. You may use these features to determine someone's gender and age, and they

also provide useful information on differentiating demographic traits. To keep with the original study methods, we use clustering algorithms to create synthetic age groups before classifying the data based on age, as explicit age labels are not accessible.

The ten acoustic features used by the framework to describe speaker-specific voice patterns are as follows: spectral centroid, first quartile (Q25), median frequency, mode, third quartile (Q75), mean dominant frequency (meandom), standard deviation (sd), and interquartile range (meanfun). As a whole, these qualities encompass variances in pitch, frequency distribution, spectral properties, and signal variability that are strongly correlated with age and gender. In order to find discriminative correlations among the chosen variables and to look at feature distributions, pairwise feature visualisation is done before model training. Prior to categorisation, the input dataset is subjected to many preparation steps. Label Encoding is used to convert the gender labels into binary number representations, and missing values are suitably handled. After that, we use StandardScaler to make sure that all the numerical characteristics are normal, meaning that their means and variances are both zero. When optimising a model, standardisation improves numerical stability and speeds convergence. In order to ensure uniformity across all tests, the processed dataset is subsequently split into 60% training data and 40% testing data.

We test two different categorisation models for gender prediction. As the first binary classifier, Logistic Regression estimates the likelihood that a given audio sample is male or female according to the chosen acoustic characteristics. Concurrently, a network with Long Short-Term Memory (LSTM) is trained to detect speech signals' temporal relationships. Two 100-hidden-unit LSTM layers make up the LSTM architecture, and to prevent overfitting, a 0.3-rate dropout layer follows. For effective training while maintaining the initial experimental setup, the Adam optimiser is used with a learning rate of 0.001 and a batch size of 32 to optimise the model.

Using the standardised acoustic properties, K-Means clustering divides the voice samples into six synthetic age groups with the following labels: 0–10, 11–20, 21–30, 31–40, 41–60, and above 60 years. Logistic Regression uses the created cluster labels as categorical targets, allowing for age categorisation with many classes. Model performance is evaluated by the use of confusion matrix analysis, Accuracy, Precision, Recall, MAE, MSE, and RMSE. An efficient framework for voice-based demographic prediction is provided by this hybrid technique, which preserves the original research contribution while combining the interpretability of traditional machine learning with the sequential learning power of deep neural networks.

IV. VOICE DATASET AND FEATURE ENGINEERING

Choosing meaningful acoustic parameters and using a high-quality speech dataset are crucial to the success of any voice-based demographic prediction system. Because of the many ways in which age and gender impact the physiological elements of speech production, it is helpful to be able to differentiate across demographic groups using carefully extracted voice parameters. To capture fluctuations in pitch, frequency distribution, and signal dynamics, the proposed system uses a structured feature engineering technique that integrates statistical and spectral descriptors. Both synthetic age group prediction and gender categorisation rely on these typical qualities.

Using recordings of both male and female speakers, the Voice Gender Dataset is used in the research. From the spoken stream, many acoustic measurements are taken to represent each recording. The dataset does not provide clear age annotations, but it does have gender designations. To overcome this shortcoming without changing the initial study approach, the system uses K-Means clustering to classify audio samples into six artificial age groups according to their shared acoustic characteristics. Predicting a person's age using this method requires no changes to the dataset or experimental setup.

From the voice signals, eleven acoustic parameters were chosen to characterise attributes particular to the speakers. Among these characteristics are the following: spectral centroid, first quartile (Q25), median frequency, mode, third quartile (Q75), standard deviation (sd), interquartile range (IQR), mean fundamental frequency (meanfun), and median frequency. The prediction models are able to pick up on modest age and gender variations thanks to these descriptors, which together reflect the statistical variability and spectral distribution of the speech signal. The chosen features' capacity to capture distinguishing voice traits while preserving computing efficiency has led to their widespread use in the field of speech processing research.
Data quality and consistent feature representation are ensured by a series of preprocessing activities carried out before to model building. To remove incomplete observations that might impact prediction accuracy, missing values are recognised and treated correctly. By using binary variables to represent the male and female classes, Label Encoding converts the category gender labels into numerical values. The next step is to use StandardScaler to standardise all numerical features. This will ensure that the feature distributions are normalised to zero mean and unit variance. By implementing standards, features with greater numerical ranges are not allowed to dominate model training, and optimisation convergence is accelerated.

Pair plot visualisations are created for both gender and synthetic age groups to further investigate the correlations among the chosen acoustic parameters. Visual representations of class-specific clustering patterns and

auditory feature pairwise distributions and correlations are provided by these studies. The visualisation shows that there is substantial divergence between male and female speakers for various parameters, especially mean fundamental frequency. This provides helpful data for future classification tasks. In addition to improving the accuracy of demographic predictions, exploratory analyses like this help find very useful traits. The original paper's pair plots (Figures 1 and 2) corroborate these findings by depicting the distribution of features for gender and age groups.

As a whole, the feature visualisation, thorough preprocessing, and carefully chosen acoustic descriptors provide a solid groundwork for the next deep learning and machine learning models. The suggested framework improves the accuracy and reliability of voice-based demographic prediction by giving strong input representations for both Logistic Regression and Long Short-Term Memory (LSTM) models, while maintaining the original feature extraction methodology and guaranteeing consistent data preparation.

V. HYBRID DEEP LEARNING PREDICTION FRAMEWORK

To forecast demographics from audio recordings, the suggested prediction framework combines supervised and unsupervised learning methods with deep learning and supervised machine learning. The framework assesses various algorithms to take advantage of their unique strengths rather than depending on a single predictive model. For improved gender recognition, Logistic Regression is used as a baseline classifier; LSTM records temporal speech features; and K-Means clustering creates artificial age groups prior to Logistic Regression's age classification. When it comes to computational efficiency, model interpretability, and predictive performance, this hybrid design strikes a good balance.

For gender prediction, the first model employs Logistic Regression, a probabilistic binary classification algorithm that estimates the likelihood of a speaker belonging to either the male or female class. During training, the classifier takes the chosen acoustic features as input variables and figures out how they relate to the binary gender labels. The dataset is divided into 60% training samples and 40% testing samples, allowing the trained model to be evaluated using unseen voice recordings. Classification performance is assessed through Accuracy, Recall (Sensitivity), Specificity, and Confusion Matrix analysis, providing a comprehensive evaluation of prediction capability.

The framework also makes use of a Long Short-Term Memory (LSTM) network to boost the accuracy of gender predictions. Unlike Logistic Regression, which assumes feature independence, LSTM processes sequential information through memory cells capable of retaining long-term temporal dependencies present within speech signals. Since variations in pitch, frequency, and vocal characteristics evolve over time, LSTM is particularly suitable for analyzing voice recordings. The network architecture consists of two LSTM layers containing 100 hidden units each, followed by a dropout layer with a dropout rate of 0.3 to minimize overfitting. Model optimization is performed using the Adam optimizer with a learning rate of 0.001 and a batch size of 32, maintaining the original experimental configuration while ensuring stable convergence during training.

For age prediction, the framework adopts a hybrid methodology combining K-Means clustering with Logistic Regression. Initially, K-Means partitions the standardized acoustic features into six synthetic age groups, representing the ranges 0–10, 11–20, 21–30, 31–40, 41–60, and above 60 years. These cluster assignments are later interpreted as categorical class labels for Logistic Regression, allowing multi-class age classification despite the lack of explicit age annotations in the original dataset. This technique respects the original research contribution while delivering a practical solution for age estimation via unsupervised learning.

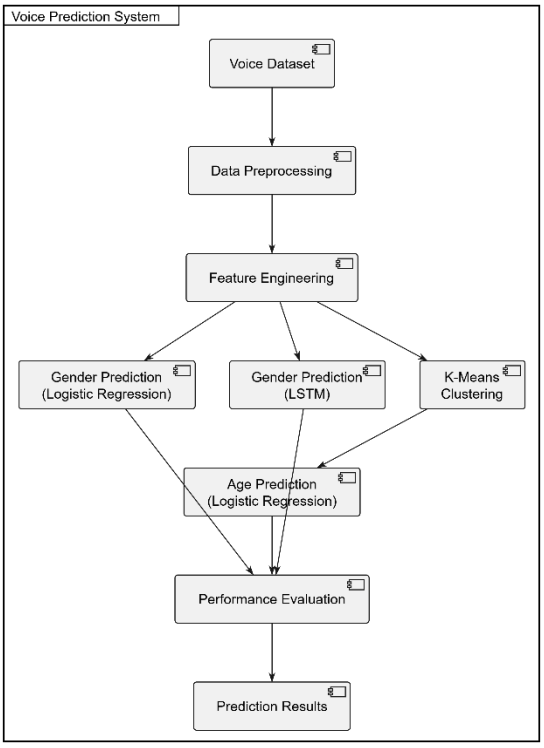

Fig. 1. Component Diagram for Hybrid Voice-Based Gender and Age Prediction Framework

The efficacy of the proposed framework is assessed using several performance indicators. Gender categorisation performance is studied using Accuracy, Precision, Recall, Specificity, and Confusion Matrix analysis, whilst age prediction is tested using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These complementing assessment measures give thorough insight into both categorisation accuracy and prediction error. Furthermore, visualization approaches,

including pair plots and confusion matrices, are applied to highlight feature distributions, class separability, and overall model performance.

By combining Logistic Regression, Long Short-Term Memory, and K-Means clustering inside a single prediction pipeline, the proposed framework effectively blends the interpretability of traditional machine learning with the sequential learning power of deep neural networks. The hybrid design successfully captures complex speech features while retaining computing efficiency, making it suited for real-world applications such as intelligent virtual assistants, healthcare systems, biometric identification, and tailored speech-based services.

VI. EXPERIMENTAL ANALYSIS

To find out how well the suggested hybrid framework might estimate age and gender from audio recordings, it was put to the test experimentally. The evaluation allows for an objective assessment of the prediction capabilities of deep learning approaches and standard machine learning algorithms by comparing them under equal experimental settings. For the sake of objectivity and consistency in the comparison analysis, all experiments used the same preprocessed dataset, chosen acoustic characteristics, and training/testing partitions.

Logistic Regression (LR), Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU), Feedforward Neural Networks (FNN), Support Vector Machines (SVM) with quadratic and Gaussian kernels, Bayesian Networks, Linear Discriminant Analysis (LDA), and K-Nearest Neighbours (KNN) were among the classification models used for gender prediction. Because of their superior capacity to grasp the intricate temporal correlations inherent in voice signals, deep learning architectures often outperform more traditional machine learning methods, as seen in the comparison. The LSTM network proved its better capacity for modelling sequential voice features with the greatest classification accuracy of 98.5% among all models that were assessed.

With a 97.7 percent accuracy rate, the baseline Logistic Regression model proved that, despite their simplistic structure, statistical machine learning methods are still quite successful for binary gender categorisation. Similarly, GRU attained 97.7 percent accuracy, while Recurrent Neural Networks and Feedforward Neural Networks both reached 97.5%. There were also competitive results from traditional classifiers; for example, Bayesian Networks achieved 95.8% accuracy, Gaussian Kernel SVM and K-Nearest Neighbours reached 97.8%, and Quadratic Kernel SVM reached 97.7%. These results show that LSTM reliably achieves the best prediction performance by efficiently learning temporal speech relationships, even though there are a number of well-performing machine learning models. The suggested hybrid method use a combination of K-Means clustering and Logistic Regression to artificially categorise speech samples into six distinct age groups for the purpose of age prediction. Since the original dataset does not include clear age labels, the standardised acoustic characteristics are first partitioned into representative age groups using K-Means clustering. Then, supervised classification is performed using Logistic Regression. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) are the metrics used to assess the performance of a model. The suggested Logistic Regression model was evaluated using Support Vector Machines, Random Forest, and Decision Trees, and it obtained MAE = 0.14, MSE = 0.66, and RMSE = 0.57. While maintaining computational simplicity, our experimental findings show that the clustering-assisted classification technique gives accurate age prediction.

Incorporating visualisation tools into the experimental study further validated the usefulness of the chosen sound elements. Visualisations of the correlations among the chosen voice descriptors for gender and age groups are shown in pair plots, which show clear grouping patterns that lend credence to demographic categorisation. Female speakers often display higher pitch values than male speakers, according to the distribution of the mean fundamental frequency (meanfun), and the confusion matrix constructed for the LSTM model accurately separates male and female voice samples. Qualitative evidence in support of the quantitative performance produced by the prediction models is provided by these visual findings. The results are further supported by the pair plots, confusion matrix, and mean fun distribution that can be seen in the graphics of the article.

In conclusion, the experimental assessment indicates that voice-based demographic prediction is much enhanced by combining deep learning with traditional machine learning. When it comes to synthetic age group categorisation, a mix of K-Means clustering and Logistic Regression yields excellent results, while the LSTM model outperforms all others when it comes to gender classification by accurately capturing the dynamics of temporal speech. It follows that the suggested hybrid architecture is a solid and extensible method for demographic forecasting from audio signals, as confirmed by the experimental results.

VII. CONCLUSION

Through the integration of Logistic Regression, Long Short-Term Memory (LSTM) networks, and K-Means clustering, this research proposes a hybrid machine learning and deep learning framework that can predict gender and age from voice recordings. The suggested approach successfully identifies speaker-specific sound patterns by integrating meticulously chosen acoustic characteristics with thorough preprocessing and feature engineering. The system offers a computationally interpretable and accurate solution for autonomous demographic prediction by combining traditional statistical learning with sequential deep learning models.

The experimental comparison shows that LSTM achieves a 98.5% accuracy rate for gender classification, which is better than the baseline Logistic Regression model. This proves that LSTM is better at learning the temporal correlations in speech signals. Though it was competitive with other traditional ML algorithms like Support Vector Machines, K-Nearest Neighbours, Bayesian Networks, and Linear Discriminant Analysis, LSTM's sequential learning capacity allowed for more precise representation of dynamic speech features. When it comes to voice-based classification tasks, where temporal information is crucial, these results show how important deep learning is.

We obtained reliable predictive performance utilising standardised auditory cues and effectively developed synthetic age categories using the recommended combination of K-Means clustering and Logistic Regression for age prediction. Extending the application of voice-based demographic prediction, the clustering-based technique shows that relevant age-related information may be recovered even without explicit age labels. Mean fundamental frequency, spectral flatness, dominant frequency, interquartile range, and spectral centroid are some of the acoustic descriptors that were chosen because they accurately portray the statistical and spectral features of speech, which in turn leads to strong classification performance.

Improving feature extraction methods that can handle noisy real-world settings, adding more advanced deep learning architectures like Transformers and attention-based recurrent networks, and expanding the suggested framework could be the focus of future research. It might be possible to improve the system's generalisability and practical implementation by researching speaker-independent models and cross-language demographic prediction. Future applications in healthcare, biometric authentication, intelligent virtual assistants, and personalised human-computer interaction can be supported by the proposed hybrid framework, which combines Logistic Regression, Long Short-Term Memory (LSTM), and K-Means clustering to provide an accurate, scalable, and dependable solution for voice-based demographic prediction.

REFERENCES

1. F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," Applied Acoustics, vol. 156, pp. 351–358, Nov. 2019.
2. S. Fahmeeda, M. Ayan, M. Shamsuddin, and A. Amreen, "Voice-based gender recognition using deep learning," International Journal of Innovative Research & Growth, vol. 3, pp. 649–654, 2022.
3. K. Chachadi and S. R. Nirmala, "Gender recognition from speech signal using 1-D CNN," in Proc. 2nd Int. Conf. Recent Trends in Machine Learning, IoT, Smart Cities and Applications (ICMISC), Singapore, 2022, pp. 349–360.
4. M. A. Uddin, R. K. Pathan, M. S. Hossain, and M. Biswas, "Gender and region detection from human voice using a three-layer feature extraction method with 1D CNN," Journal of Information and Telecommunication, vol. 6, no. 1, pp. 27–42, 2022.
5. A. A. Alnuaim et al., "Speaker gender recognition based on deep neural networks and ResNet50," Wireless Communications and Mobile Computing, vol. 2022, Art. no. 4444388, 2022.
6. A. Ksibi et al., "Voice pathology detection using a two-level classifier based on combined CNN–RNN architecture," Sustainability, vol. 15, no. 4, p. 3204, 2023.
7. C. Adhithi, N. Chandana, B. Nikita, and D. Shravani, "Gender recognition using voice," International Journal of Multidisciplinary Research, 2023.
8. A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, "Age and gender recognition using a convolutional neural network with a multi-attention module through speech spectrograms," Sensors, vol. 21, no. 17, p. 5892, 2021.
9. Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks," IEEE Access, vol. 8, pp. 16560–16572, 2020.
10. M. Alsulaiman, Z. Ali, and G. Muhammad, "Gender classification with voice intensity," in Proc. UKSim European Symposium on Computer Modeling and Simulation, 2011, pp. 205–209.
11. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
12. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
13. C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
14. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.
15. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2023.
16. S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.
17. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 1967, pp. 281–297.
18. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. International Conference on Learning Representations (ICLR), 2015.

19. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
20. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
21. A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6645–6649.
22. K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in Proc. EMNLP, 2014, pp. 1724–1734.
23. A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.
24. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
25. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.