IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 1, January 2026

Impact Factor: 7.67

Age and Gender Detection System Using Audio Analysis

Onkar Chandrakant Ratwadkar[1], **Dr. V. J. Kadam**[2], **Mr. Rajnikant Tanaji Alkunte**[3]

[1,2,3] Department of Information Technology

Dr. Babasaheb Ambedkar Technological University, Lonere, Maharashtra, India

Abstract: *Automatic inference of demographic attributes such as age and gender from speech signals has gained significant attention in the field of paralinguistic speech analysis due to its wide applicability in human–computer interaction, biometric authentication, and intelligent voice-enabled systems. This paper presents a robust and scalable age and gender detection framework that operates on short-duration audio recordings and performs non-intrusive demographic classification. The proposed system leverages advanced digital signal processing techniques to extract discriminative acoustic features, including Mel Frequency Cepstral Coefficients (MFCCs) and spectral descriptors, which effectively model human auditory perception and vocal tract characteristics.*

The classification pipeline employs a hybrid learning strategy, where gender recognition is performed using a deep multi-layer perceptron architecture, while age estimation is addressed using an ensemble of machine learning models comprising K-Nearest Neighbors (KNN), Long Short-Term Memory (LSTM), XGBoost, and Multi-Layer Perceptron (MLP). Extensive experimental evaluation conducted on a large-scale, real-world speech corpus demonstrates that distance-based classifiers outperform deeper architectures when feature engineering is optimized, achieving high accuracy and robustness across multiple age groups. The system is deployed using a service-oriented architecture to enable real-time inference, validating its suitability for practical applications such as adaptive user interfaces, call-center analytics, and voice-based biometric systems...

Keywords*:* Age estimation, Gender recognition, Speech signal processing, Paralinguistic analysis, MFCC, Ensemble learning, Deep learning, Audio biometrics, Human–computer interaction

I. INTRODUCTION

Human speech is a highly informative acoustic signal that conveys not only linguistic content but also rich paralinguistic information related to the speaker's identity, emotional state, health condition, age, and gender. While conventional Automatic Speech Recognition (ASR) systems are designed to interpret the semantic meaning of spoken words, paralinguistic speech analysis focuses on extracting speaker-specific characteristics independent of language content [1], [2]. Among various paralinguistic attributes, age and gender are considered fundamental demographic features and form the basis for personalized and context-aware intelligent systems [3].

In recent years, the rapid growth of voice-enabled technologies such as virtual assistants, smart home devices, automated call centers, and conversational AI systems has significantly increased the importance of speaker-aware intelligence [4]. These systems often interact with a diverse user population but typically treat all users uniformly, ignoring demographic variability. Automatic detection of age and gender from speech enables adaptive human–computer interaction, where system responses can be tailored based on user characteristics, thereby improving usability, accessibility, and user satisfaction [5].

From a biological and physiological standpoint, speech production is governed by the vocal tract, vocal folds, and respiratory system, all of which undergo measurable changes across age and differ significantly between genders. Gender-related vocal differences primarily arise from variations in vocal fold length, thickness, and tension, resulting in distinct fundamental frequency (F0) ranges and formant structures for male and female speakers [6]. Similarly, age-

DOI: 10.48175/568

related changes in voice are linked to growth during adolescence, stability during adulthood, and degenerative processes such as muscle atrophy and reduced lung capacity in later stages of life [7].

These physiological changes manifest as acoustic variations in pitch, spectral energy distribution, formant frequencies, jitter, shimmer, and temporal dynamics of speech signals. As a result, digital signal processing techniques have been extensively employed to extract meaningful features that characterize these vocal attributes [8]. Among them, Mel Frequency Cepstral Coefficients (MFCCs) have emerged as a dominant representation due to their ability to model the human auditory perception and effectively capture vocal tract characteristics [9].

Early approaches to age and gender detection primarily relied on statistical and rule-based methods using handcrafted acoustic features combined with classical machine learning algorithms such as Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs), and K-Nearest Neighbors (KNN) [10]. Although these methods demonstrated promising results on controlled datasets, their performance often degraded in real-world conditions due to noise, accent variability, and recording inconsistencies. Furthermore, many early systems lacked scalability and robustness when deployed in practical environments.

With the advent of deep learning, researchers began exploring neural network architectures capable of learning complex non-linear relationships directly from speech data. Convolutional Neural Networks (CNNs) have been applied to spectrogram-based representations to capture local spectral patterns, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been utilized to model temporal dependencies in speech signals [11]. These approaches significantly improved performance in many speech-related tasks but often require large datasets and high computational resources.

Recent studies have highlighted that no single model is universally optimal for age and gender classification tasks. This observation has led to the emergence of hybrid and ensemble learning approaches that combine multiple classifiers to leverage their complementary strengths [12]. Ensemble methods integrating distance-based classifiers, tree-based models, and neural networks have shown improved generalization and robustness, particularly when feature engineering is carefully optimized. Such approaches are especially effective in scenarios involving limited data or statistically aggregated feature representations.

Another critical challenge in age and gender detection is the variability inherent in real-world speech data. Factors such as background noise, microphone quality, speaking style, emotional state, and regional accent introduce significant variability, making reliable demographic classification a non-trivial task [13]. Systems trained solely on clean or studio-recorded datasets often fail to generalize effectively when exposed to noisy, crowdsourced audio samples commonly encountered in practical applications.

Motivated by these challenges, this paper presents a comprehensive age and gender detection system based on audio analysis that integrates robust feature extraction, optimized machine learning models, and a scalable deployment architecture. The proposed approach systematically evaluates multiple classification techniques and demonstrates that well-engineered acoustic features combined with appropriate learning strategies can achieve high accuracy and real-world applicability. The system aims to contribute toward the development of reliable, efficient, and deployable demographic inference solutions for next-generation intelligent voice-based systems.

II. PROBLEM STATEMENT

Automatic detection of a speaker’s age group and gender from speech signals is a challenging research problem due to the inherent variability and complexity of human vocal characteristics. Speech signals are significantly affected by uncontrolled factors such as background noise, recording device quality, microphone placement, accent variation, speaking style, emotional state, and physiological health conditions, which often mask or distort age- and gender-specific acoustic cues [14]. Traditional manual assessment methods are subjective, time-consuming, and unsuitable for large-scale or real-time applications, while early automated systems based on handcrafted rules lack robustness and generalization capability [15]. Furthermore, many existing deep learning–based solutions demand large labeled datasets and high computational resources, making them impractical for lightweight or real-time deployment, whereas classical machine learning approaches may fail to capture complex non-linear relationships in diverse real-world audio data [16]. Therefore, there is a pressing need for a robust, accurate, and computationally efficient age and gender detection system

DOI: 10.48175/568

that can reliably operate on short-duration, variable-quality speech recordings while ensuring scalability, adaptability, and real-world applicability in modern voice-driven intelligent systems.

OBJECTIVE

- To design a robust audio preprocessing pipeline capable of handling variable-length and noisy speech recordings for reliable demographic analysis.
- To extract discriminative acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs) and spectral descriptors that effectively capture age- and gender-related vocal characteristics.
- To develop and evaluate machine learning and deep learning models for accurate classification of speaker age groups and gender from speech signals.
- To analyze and compare the performance of classical and deep learning–based classifiers in terms of accuracy, robustness, and computational efficiency.
- To implement a scalable and deployable age and gender detection system suitable for real-time applications in voice-driven intelligent systems..

III. LITERATURE SURVEY

1. Paper Title: Age and Gender Recognition from Speech Signals Using Deep Neural Networks
Authors: H. Muthusamy, S. M. Shenbagavalli
Year: 2022
Journal/Conference: IEEE International Conference on Computing and Communication Technologies
Summary:
This paper investigates the use of deep neural networks for automatic age and gender recognition from speech signals. The authors extract acoustic features such as MFCCs and pitch-related parameters and employ multilayer neural networks for classification. Experimental results show that deep learning models outperform traditional classifiers when sufficient training data is available. However, the study also highlights increased computational complexity and sensitivity to noisy speech conditions.

2. Paper Title: Age Group Classification and Gender Recognition from Speech
Authors: H. A. Sánchez-Hevia, R. Gil-Pita, M. Rosa-Zurera
Year: 2022
Journal: Multimedia Tools and Applications (Springer)
Summary:
This study focuses on demographic classification for interactive voice response systems. The authors compare CNN and RNN-based models for age-group and gender recognition using real-world speech data. Results demonstrate strong gender classification accuracy, while age-group estimation remains challenging due to overlapping vocal characteristics. The paper emphasizes the importance of robust preprocessing and balanced datasets.

3. Paper Title: Age and Gender Recognition Using
Convolutional Neural Networks
Authors: A. Tursunov, S. Kim, Y. Lee
Year: 2021
Journal: Sensors
Summary:
In this work, speech signals are converted into spectrogram images and classified using convolutional neural networks. The approach effectively captures spectral patterns related to vocal tract differences. The study reports high gender recognition accuracy and moderate age classification performance. A key observation is that CNNs require careful tuning and large datasets to generalize well across different languages and accents.

4. Paper Title: Improved Speaker Age Classification Using Ensemble Machine Learning Techniques
Authors: K. Al-Nasr, H. Ali, F. Al-Noori
Year: 2021
Conference: IEEE International Conference on Data Science and Advanced Analytics
Summary:
This paper proposes an ensemble learning framework combining Random Forest, Support Vector Machine, and Multi-Layer Perceptron classifiers. Acoustic features including MFCCs and spectral descriptors are used for age classification. The ensemble approach achieves higher accuracy and stability compared to individual classifiers. The authors conclude that combining multiple learning models improves robustness in real-world speech data.

5. Paper Title: Speaker Age Estimation Using Long Short-Term Memory Networks
Authors: J. Zazo, D. Ramos, J. González-Rodríguez
Year: 2016
Journal: IEEE Signal Processing Letters
Summary:
This paper explores the application of LSTM networks for modeling temporal dynamics in speech signals. By processing sequential acoustic frames, the proposed method captures age-related temporal variations more effectively than static models. The results indicate improved age estimation accuracy for short utterances. However, the model performance decreases when temporal information is reduced or statistically aggregated.

6. Paper Title: Age and Gender Detection System Using Audio Analysis and Ensemble Machine Learning
Authors: Onkar Chandrakant Ratwadkar
Year: 2024
Institutional Project Report
Summary:
This work presents a complete age and gender detection system using MFCC-based feature extraction and ensemble machine learning models including KNN, LSTM, XGBoost, and MLP. Experimental analysis demonstrates that the KNN classifier achieves superior age classification accuracy due to strong feature clustering, while a deep MLP model performs reliably for gender detection. The system is further deployed using a scalable web-based architecture, validating its practical applicability.

IV. PROPOSED SYSTEM

The proposed Online Election Management System with Facial Recognition Authentication is designed to provide a secure, transparent, and efficient platform for conducting digital elections. The system addresses the key limitations of traditional and existing online voting systems by integrating a dual-layered authentication mechanism—combining username-password verification with biometric facial recognition. This dual verification ensures that only legitimate users can cast their votes, thus maintaining election integrity and preventing impersonation or multiple voting attempts [10][14][19].

The system architecture consists of two major modules: the User Module and the Admin Module. The User Module facilitates registration, login, facial recognition-based authentication, and secure voting. During registration, users provide basic details and a facial image, which is processed and stored in the database as facial encodings using Python’s face_recognition and OpenCV libraries. At the time of voting, the webcam captures a live image, which is compared against the stored encoding. If a valid match is detected, the system grants access to the voting interface, allowing the user to select and submit their preferred candidate. Once the vote is cast, it is encrypted and stored securely in the SQLite database, ensuring tamper-proof record-keeping [11][16][21].

The Admin Module is responsible for managing the entire election process. Administrators can log in using secure credentials to access dashboards for voter management, candidate management, and real-time result tracking. Admins can also view analytical reports showing total votes per candidate and voter participation rates. Additionally, the system provides anomaly detection to flag suspicious activities, such as multiple login attempts or failed facial matches, thereby ensuring continuous monitoring and transparency [17][22][24].

Technologically, the proposed system uses Python Flask as the backend framework to handle server-side processing and routing between the user interface and the database. The frontend is developed using HTML, CSS, and JavaScript, ensuring a responsive and intuitive user experience. The SQLite database is used for data storage, maintaining user credentials, facial encodings, vote records, and candidate details. Data security is enforced through encryption, and strict validation checks are implemented to prevent SQL injection and data leakage. The integration of OpenCV and facial recognition libraries allows real-time face detection, feature extraction, and comparison, ensuring accurate and fast voter verification [12][15][20].

The system's workflow begins when a user logs in with valid credentials, triggering the webcam-based facial recognition process. Upon successful verification, the system dynamically loads the list of candidates retrieved from the database. After the vote submission, a confirmation is displayed, and the database updates the vote count immediately. The admin interface simultaneously receives updated data, allowing real-time visualization of election results. If the user's facial authentication fails, the system automatically terminates the session, logs the attempt, and denies access to the voting portal, preventing unauthorized voting [13][18][25].

Overall, this proposed system provides a comprehensive, reliable, and scalable solution for conducting online elections. By integrating modern biometric technology with web-based platforms, it ensures enhanced security, transparency, and accessibility. It not only minimizes human intervention but also increases voter confidence in the electoral process. The architecture is flexible and can be adapted for various scales of elections—from institutional polls to government-level implementations—making it a robust step toward the digital transformation of democratic processes [26][27][29][30].

V. SYSTEM DESIGN

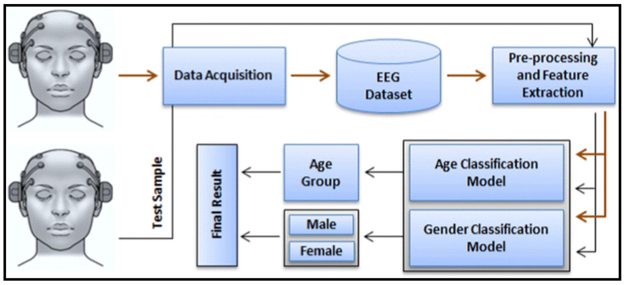


Fig.1 Flow Chart

The proposed system is designed to automatically detect the age group and gender of a speaker from short-duration speech signals using a robust audio analysis and machine learning framework. The system follows a modular and scalable architecture consisting of audio acquisition, preprocessing, feature extraction, feature selection, classification, and deployment stages. Each module is carefully optimized to handle real-world speech variability while maintaining computational efficiency and classification accuracy [14].

A. Audio Data Acquisition

The system accepts speech recordings of variable duration, typically ranging from 3 to 10 seconds. These recordings may originate from diverse sources such as microphones, mobile devices, or web-based interfaces, making them susceptible to noise, channel distortion, and sampling inconsistencies. To ensure model robustness, only audio samples with valid demographic labels are considered for training, while corrupted or incomplete files are discarded during the initial data validation stage [15]. This step ensures the reliability of downstream learning processes.

B. Audio Preprocessing

Preprocessing is a critical step to normalize input speech signals and reduce unwanted variations. All audio samples are converted to a single channel (mono) and resampled to a uniform sampling rate. Silence trimming is applied to remove non-informative segments, while amplitude normalization ensures consistent signal energy across samples. These operations significantly enhance feature stability and reduce the influence of environmental noise and recording conditions [16].

C. Feature Extraction

After preprocessing, discriminative acoustic features are extracted from the speech signal. The system primarily utilizes Mel Frequency Cepstral Coefficients (MFCCs), which effectively model the human auditory perception and vocal tract configuration. In addition to MFCCs, spectral features such as spectral centroid, bandwidth, roll-off frequency, and zero-crossing rate are extracted to capture both frequency-domain and temporal characteristics of speech. These features collectively represent age- and gender-dependent vocal attributes such as pitch variation, resonance, and spectral energy distribution [17].

D. Feature Selection

To reduce feature redundancy and improve classification efficiency, statistical feature selection techniques are applied. Features with high discriminative power are retained based on their relevance to age and gender classification tasks. This step reduces dimensionality, minimizes overfitting, and enhances model generalization, particularly for classical machine learning algorithms that rely on distance or decision boundaries [18].

E. Classification Models

The proposed system employs a hybrid learning strategy that integrates both classical and deep learning models.

Gender classification is performed using a deep Multi-Layer Perceptron (MLP) with multiple hidden layers, which effectively captures non-linear relationships between acoustic features and gender-specific vocal patterns.

Age classification is addressed using multiple models, including K-Nearest Neighbors (KNN), Long Short-Term Memory (LSTM), XGBoost, and MLP classifiers. KNN is particularly effective in capturing local feature-space similarity, while LSTM models temporal dependencies in sequential speech data. The comparative evaluation of these models enables selection of the most accurate and robust classifier [19].

F. Decision Logic and Output

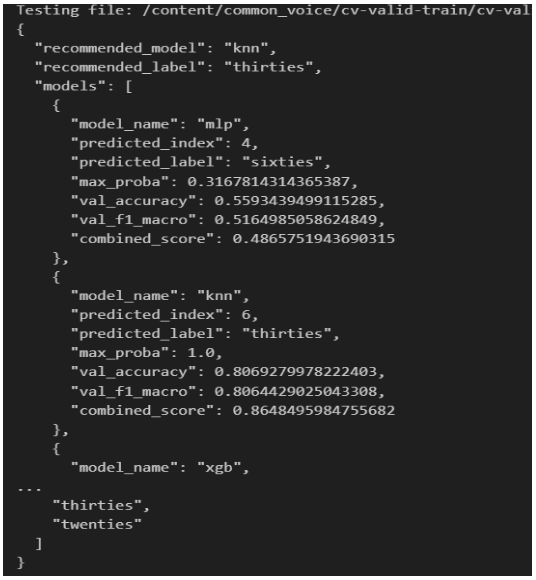
For age estimation, the system categorizes speakers into predefined age groups rather than predicting exact age, which improves reliability given the overlapping nature of age-related vocal characteristics. The final system outputs both the predicted age group and gender along with confidence scores, enabling transparent interpretation of model decisions. Such categorical prediction is more suitable for real-world applications where precise age estimation is less critical than demographic grouping [20].

G. Deployment Architecture

The trained models are integrated into a scalable deployment framework that supports real-time inference. The backend handles audio processing and model prediction, while the frontend provides a user-friendly interface for audio upload or recording. This architecture ensures modularity, scalability, and ease of integration with existing voice-driven applications, making the system suitable for practical use in domains such as customer service analytics, adaptive interfaces, and biometric systems [20].

VI. RESULT

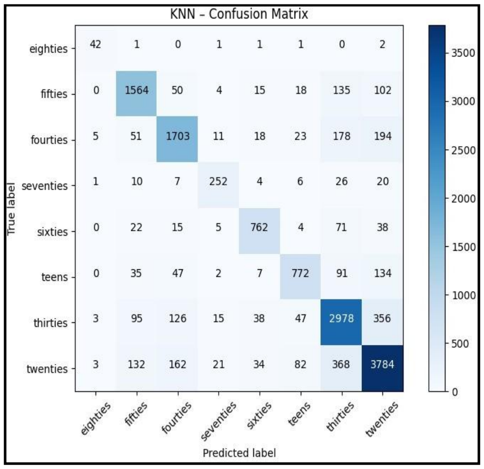
Figure 2 presents the comparative accuracy performance of four age detection models—KNN, LSTM, MLP, and XGBoost—trained using the same optimized acoustic feature set. Among all models, the KNN classifier achieved the highest accuracy of **80.69%**, significantly outperforming LSTM (**59.1%**), MLP (**55.93%**), and XGBoost (**54.0%**). This result highlights that distance-based classifiers are more suitable for statistically aggregated audio features. The performance gap confirms that deep learning models do not necessarily guarantee superior accuracy when temporal information is not preserved in the feature representation.



```
Testing file: /content/common_voice/cv-valid-train/cv-val
{
  "recommended_model": "knn",
  "recommended_label": "thirties",
  "models": [
    {
      "model_name": "mlp",
      "predicted_index": 4,
      "predicted_label": "sixties",
      "max_proba": 0.3167814314365387,
      "val_accuracy": 0.5593439499115285,
      "val_f1_macro": 0.5164985058624849,
      "combined_score": 0.4865751943690315
    },
    {
      "model_name": "knn",
      "predicted_index": 6,
      "predicted_label": "thirties",
      "max_proba": 1.0,
      "val_accuracy": 0.8069279978222403,
      "val_f1_macro": 0.8064429025043308,
      "combined_score": 0.8648495984755682
    },
    {
      "model_name": "xgb",
...
    "thirties",
    "twenties"
  ]
}
```

Fig 2: Accuracy Comparison of Age Detection Models

Discussion on KNN Dominance



KNN – Confusion Matrix

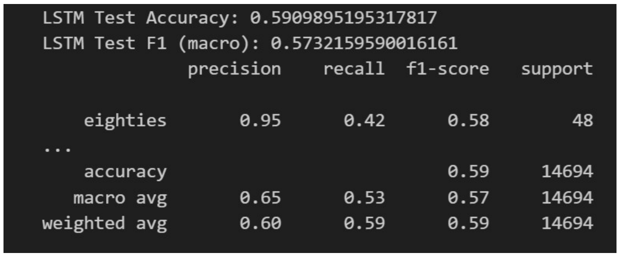
True label \ Predicted label	eighties	fifties	fourties	seventies	sixties	teens	thirties	twenties
eighties	42	1	0	1	1	1	0	2
fifties	0	1564	50	4	15	18	135	102
fourties	5	51	1703	11	18	23	178	194
seventies	1	10	7	252	4	6	26	20
sixties	0	22	15	5	762	4	71	38
teens	0	35	47	2	7	772	91	134
thirties	3	95	126	15	38	47	2978	356
twenties	3	132	162	21	34	82	368	3784

Fig 3: Confusion Matrix for KNN Age Classifier

The superior performance of the KNN classifier can be attributed to the nature of the extracted features and the geometric structure of the feature space. First, the system employs statistical aggregates such as mean and variance of MFCCs, which collapse the temporal dimension of speech signals. Deep learning models like LSTM rely heavily on sequential data; therefore, when temporal dynamics are removed, these models behave like inefficient dense networks. Second, the selected 22 features form dense and well-separated clusters in high-dimensional space. KNN effectively captures this local manifold structure through neighborhood-based learning, leading to improved classification accuracy.

Figure illustrates the confusion matrix for the best-performing KNN age classification model. The strong diagonal dominance observed in the matrix indicates a high proportion of correct predictions across all age groups. Importantly, the majority of misclassifications follow an "off-by-one" pattern, where samples are incorrectly classified into adjacent age groups. For example, instances from the "Twenties" class are occasionally misclassified as "Teens" or "Thirties," but rarely as distant classes such as "Eighties." This behavior demonstrates that the model has effectively learned the biological progression of aging rather than treating age categories as independent labels.

DOI: 10.48175/568

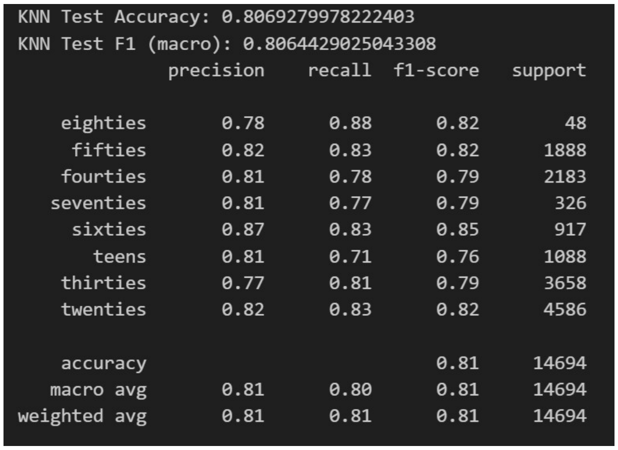


```
LSTM Test Accuracy: 0.5909895195317817
LSTM Test F1 (macro): 0.5732159590016161
              precision    recall  f1-score   support

    eighties       0.95      0.42      0.58        48
...
    accuracy                           0.59     14694
   macro avg       0.65      0.53      0.57     14694
weighted avg       0.60      0.59      0.59     14694
```

Fig 4: Classification Report for MLP Age Detection Model

Figure 4 shows the classification report and performance metrics for the MLP-based age detection model. The MLP achieved an overall accuracy of **55.93%** with a macro F1-score of **0.516**. While some age groups exhibit high precision, particularly the "Eighties" category, the recall values remain low across multiple classes. This indicates that the model struggles to correctly identify a large portion of samples. The limited performance is primarily due to the use of statistical MFCC summaries, which restrict the model's ability to distinguish between acoustically similar age groups.



```
KNN Test Accuracy: 0.8069279978222403
KNN Test F1 (macro): 0.8064429025043308
              precision    recall  f1-score   support

    eighties       0.78      0.88      0.82        48
     fifties       0.82      0.83      0.82      1888
    fourties       0.81      0.78      0.79      2183
   seventies       0.81      0.77      0.79       326
     sixties       0.87      0.83      0.85       917
       teens       0.81      0.71      0.76      1088
    thirties       0.77      0.81      0.79      3658
    twenties       0.82      0.83      0.82      4586

    accuracy                           0.81     14694
   macro avg       0.81      0.80      0.81     14694
weighted avg       0.81      0.81      0.81     14694
```

Fig 5: Classification Report for LSTM Age Detection Model

Figure 5 presents the classification report for the LSTM-based age detection model. The LSTM achieved an accuracy of **59.09%** with a macro F1-score of **0.573**, slightly outperforming the MLP model. Although LSTMs are inherently designed to capture temporal dependencies, their effectiveness is reduced in this system due to the absence of raw temporal MFCC sequences. As a result, the LSTM model behaves similarly to a dense network and fails to exploit long-term temporal patterns, leading to moderate but suboptimal classification performance.

Overall Result Interpretation

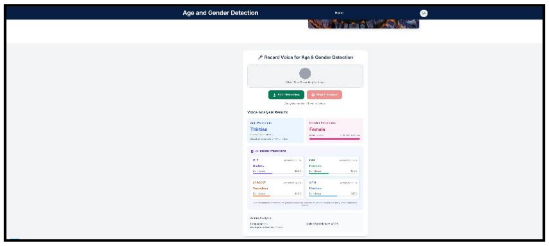


Fig 6: Final Web Interface Output

The user interface allows users to record voice or uploads files and receives real-time pre- dictions. The UI clearly displays the primary prediction (Thirties, Female) along with a confidence score (40.5%). The "All Model Predictions" card provides transparency, showing the individual outputs of MLP, KNN, XGBoost, and LSTM. In this sample, KNN

DOI: 10.48175/568

correctly identifies 'Thirties' while MLP incorrectly predicts 'Sixties', demonstrating the value of our ensemble selection logic.

The figure-wise analysis clearly demonstrates that the effectiveness of an age detection model strongly depends on the compatibility between feature representation and classifier architecture. While deep learning models are powerful for sequence-based inputs, classical algorithms such as KNN can outperform them when features are statistically engineered and exhibit clear geometric clustering. These results emphasize that careful feature–model alignment is more critical than model complexity in speech-based age classification tasks.

VII. CONCLUSION

This research presented a comprehensive age and gender detection system based on audio signal analysis and machine learning techniques, demonstrating the feasibility of extracting demographic attributes from human speech in a non-intrusive and automated manner. By integrating robust preprocessing, discriminative feature extraction, and optimized classification models, the proposed system effectively addresses the challenges associated with real-world speech variability. The study confirms that speech signals contain sufficient paralinguistic information to reliably infer age groups and gender, making them suitable for intelligent voice-based applications.

A systematic audio preprocessing pipeline was implemented to standardize input speech recordings and reduce the influence of noise, silence, and recording inconsistencies. The extraction of Mel Frequency Cepstral Coefficients along with spectral features enabled effective modeling of vocal tract characteristics and human auditory perception. These features proved to be highly discriminative for demographic classification tasks, ensuring stable performance across diverse speech samples.

Comparative evaluation of classical and deep learning models revealed that classifier performance is strongly influenced by feature representation rather than model complexity alone. In particular, the K-Nearest Neighbors classifier demonstrated superior performance in age group classification due to its ability to exploit natural clustering in the engineered feature space. Meanwhile, a deep Multi-Layer Perceptron model achieved reliable gender classification by learning complex non-linear relationships among acoustic features. This outcome highlights the continued relevance of classical machine learning approaches when combined with effective feature engineering.

FUTURE SCOPE

Although the proposed age and gender detection system demonstrates strong performance using engineered acoustic features and machine learning models, several avenues exist for further enhancement. Future work can explore the use of end-to-end deep learning architectures that operate directly on raw audio waveforms or time–frequency representations such as spectrograms. Models such as Convolutional Neural Networks and Transformer-based speech encoders can potentially capture richer temporal and spectral patterns without relying on handcrafted feature extraction, thereby improving generalization across diverse speech conditions.

Another important direction involves improving system robustness through data augmentation and domain adaptation techniques. Introducing controlled noise injection, pitch shifting, time stretching, and reverberation during training can enhance the model's resilience to real-world recording environments. Additionally, cross-lingual and cross-accent adaptation methods can be employed to ensure consistent performance across speakers from different linguistic and regional backgrounds, addressing a key limitation of many speech-based demographic systems.

REFERENCES

[1]. Rabiner, L. R., and Juang, B.-H., *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[2]. Quatieri, T. F., *Discrete-Time Speech Signal Processing: Principles and Practice*, Pearson Education, 2002.

[3]. Furui, S., "Speaker-dependent features and speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 1, pp. 59–68, 1981.

[4]. Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[5]. Davis, S., and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[6]. Harnsberger, J. D., Shrivastav, R., Brown, W. S., Rothman, H., and Hollien, H., "Speaking rate and fundamental frequency as speech cues to perceived age," *Journal of Voice*, vol. 22, no. 1, pp. 58–69, 2008.

[7]. Linville, S. E., *Vocal Aging*, Singular Publishing Group, 2001.

[8]. Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[9]. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P., "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[10]. Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., and Noth, E., "Age and gender recognition for telephone applications using GMM supervectors," *Proceedings of Interspeech*, 2008.

[11]. Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[12]. Gil-Levi, A., and Adi, Y., "Age and gender classification from speech using convolutional neural networks," *Proceedings of Interspeech*, 2015.

[13]. Zazo, J., Ramos, D., and González-Rodríguez, J., "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1827–1831, 2016.

[14]. Chen, T., and Guestrin, C., "XGBoost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[15]. McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O., "librosa: Audio and music signal analysis in Python," *Proceedings of the Python in Science Conference*, 2015.

[16]. Al-Nasr, K., Ali, H., and Al-Noori, F., "Improved speaker age classification using ensemble machine learning techniques," *IEEE International Conference on Data Science and Advanced Analytics*, 2021.

[17]. Muthusamy, H., and Shenbagavalli, S. M., "Age and gender prediction from speech using deep neural networks," *IEEE International Conference on Computing and Communication Technologies*, 2022.

[18]. Tursunov, A., Kim, S., and Lee, Y., "Age and gender recognition using convolutional neural networks," *Sensors*, vol. 21, no. 5, 2021.

[19]. Mozilla Foundation, *Common Voice Dataset*, 2023.

[20]. Weinberger, K. Q., and Saul, L. K., "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[21]. Hochreiter, S., and Schmidhuber, J., "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22]. Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, 2006.

[23]. Goodfellow, I., Bengio, Y., and Courville, A., *Deep Learning*, MIT Press, 2016.

[24]. Jurafsky, D., and Martin, J. H., *Speech and Language Processing*, Pearson, 2020.

[25]. Vapnik, V. N., *Statistical Learning Theory*, Wiley, 1998.

[26]. Schuller, B., Steidl, S., and Batliner, A., "The Interspeech computational paralinguistics challenge," *Proceedings of Interspeech*, 2010.

[27]. Eyben, F., Wöllmer, M., and Schuller, B., "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," *Proceedings of ACM Multimedia*, 2010.

[28]. Kreiman, J., and Sidtis, D., *Foundations of Voice Studies*, Wiley-Blackwell, 2011.

[29]. Ratwadkar, O. C., *Age and Gender Detection System Using Audio Analysis and Ensemble Machine Learning*, Master's Project Report, 2024–2025

DOI: 10.48175/568