INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Audio-based Age, Gender & Emotion detection

M Swapna[1]*, Chittepu Sireesha[2]*

1 Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana-500001, India
2 Department of Information Technology, Vasavi College of Engineering, Ibrahimbagh, Hyderanad, Telangana-500031, India

OPEN ACCESS

Received: 03/06/2025

Accepted: 02/09/2025

Published: 20/10/2025

Citation: Swapna M, Sireesha C (2025) Audio-based Age, Gender & Emotion detection. Indian Journal of Science and Technology 18(37): 3046-3060. https://doi.org/10.17485/IJST/v18i37.1039

* **Corresponding authors**.

mswapna@stanley.edu.in

sireesha@staff.vce.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The purpose of this study is to produce an automated system using deep learning that can guess the gender, age, and mood of a speaker from their speech. The purpose is to connect people to computers better, support early diagnostics in medicine, and make automated customer service more caring by responding in a flexible way using a person's speech. **Methods:** Three specialised deep learning models will be applied to the study: an LSTM-based classifier predicting gender, a CNN-based model classifying age based on the Mozilla Common Voice audio files, and a fine-tuned Deepsight model to predict the emotion, based on a mixture of CREMA-D, RAVDESS, SAVEE, and TESS data sets. Such data provides speech at various genders and ages with different emotions that include happiness, sadness, anger, and neutrality. Normalisation of the audio levels, noise removal, and the extraction of features on the audio data with the help of Mel-Frequency Cepstral Coefficients (MFCCs) were performed as preprocessing steps. Cross-validation was used to evaluate the models and test on unseen speech data as a way of measuring generalizability. **Findings:** The proposed system attained a gender predictive accuracy of 96.3 and an F1-score of 0.95. To classify speaking age into 10-year age groupings, the model ranked with an accuracy of 87.4%. The main emotional states were identified by the emotion recognition module with an accuracy of 91.2 per cent. K-fold cross-validation confirmed the performance, and the performance was also re-confirmed on practical audio samples. The combined pipeline model showed improved consistency and reduced computation complexity in comparison to the methods used when every model was run separately. **Novelty:** Instead of choosing one factor at a time, the present work looks at gender, age, and emotions all at once from a single voice sample. Since the end-to-end model pipeline is simple and time-saving, it helps build real-time virtual assistants, telemedicine tools, and programs focused on feelings.

Keywords: Speech Processing; Deep Learning; LSTM; Voice Recognition; Multimodal Voice Profiling

1 Introduction

The current developments in the field of artificial intelligence (AI) and deep learning have brought a considerable improvement in the human-computer interaction, namely, speech-based communication as the most natural and intuitive form of human expression. Current AI systems [1] can do more than convert speech to text; they can use voice signals to obtain soft biometric features, including gender, age, and emotional state. Numerous types of researches [2], [3] have been conducted on speech-based profiling regarding different aspects. As an example, a multi-task deep neural network identity, gender, age, and emotion recognition proposed by Foggia et al. demonstrated the promise of deep learning in soft biometrics [4]. Likewise, Jaid and Karim proposed a 1D CNN with self-attention mechanism on speaker profiling, who concentrated on speech sound embeddings via Wav2Vec2.0 [5]. Audio-based anger detection has also been done using other methods such as the hybrid ANN and fuzzy logic system proposed by Surana et al. on the CREMA-D dataset [6]. Nevertheless, these models have shown great promises; however, they usually focus on one task or propose many independent systems as a requirement per classification goal. A significant literature gap is that there are no coherent systems that are able to predict gender, age, and emotion simultaneously using only one audio input. The majority of current solutions follow an individual model per task, thereby not only making computation more complex, but also creating integration problems and real-time latency problems [7]. In addition, most conventional systems focus on the lexical characteristics and ignore paralinguistic features, including pitch, timbre, and prosody, which play a critical role in emotion and demographic recognition. [8] Besides this, there is also a problem of robustness which is caused by differences in accents, background noise and speaker-dependent vocal peculiarities which makes generalization a major problem [9].

Fig 1. Speech-Based Attribute Detection

In order to overcome such drawbacks, this study suggests a universal deep learning model of audio gender, age, and emotion recognition. The system comprises three optimized models, an LSTM network, gender classification model to efficiently extract the sequential vocal characteristics Support Vector Classifier (SVC) and Random Forest, structured age group prediction model, and CNN architecture emotion classification model based on Mel-frequency cepstral coefficients (MFCCs) and spectrogram representations. Noise reduction, normalization, silence trimming, and data augmentation are applied to preprocess the data and guarantee high-quality features extraction and models robustness. The open datasets like Mozilla Common Voice to predict age and gender and CREMA-D to classify emotions offer a realistic and broad range of training data to the models [10]. The proposed method, unlike others, would combine all tasks in one pipeline without compromising too much on accuracy or real-time capability. It can be applied in the sphere where adaptive and sensitive AI reactions are needed: telemedicine, customer interaction platforms, smart surveillance. The work belongs to the rapidly developing area of affective computing and offers a scaleable, accurate and integrated solution to the problem of audio-based speaker profiling. The suggested system illustrates how integrating task-based models into a common system can break the barrier of the fragmented design, enhance the effectiveness of processing, and provide more exciting levels of user interaction.

- **Task Fragmentation** – Most existing systems are designed for single-task learning, i.e., one model for gender, another for age, and yet another for emotion recognition. This leads to redundant computation, lack of integration, and significant latency when deployed in real-time scenarios.
- **Feature Utilization** – Many conventional models primarily rely on lexical features derived from transcribed speech, neglecting paralinguistic cues such as pitch, timbre, rhythm, and prosody. These non-verbal characteristics are crucial for robust demographic and emotional inference.
- **Generalization and Robustness** – Speech signals are highly variable due to accent diversity, speaker idiosyncrasies, background noise, and recording conditions. This variability reduces the robustness and generalizability of existing models, especially when trained on constrained datasets.
- **Model Efficiency for Real-Time Deployment** – Deep models with multiple independent subsystems are computationally heavy, which limits their scalability and usability in real-world, low-latency applications.

2 Related Work

Research in speech-based soft biometric profiling has expanded significantly in recent years, with diverse approaches targeting gender, age, and emotion recognition. Foggia et al. (2024)[(1)] pioneered the use of multi-task neural networks for simultaneous identity, gender, age, and emotion recognition, demonstrating the effectiveness of deep learning in cognitive robotics. Jaid and Abdulhasan (2024)[(2)] advanced this line of work by proposing a 1D CNN with self-attention and Wav2Vec2.0 embeddings, focusing on age and gender detection. In parallel, Surana et al. (2024)[(3)] addressed emotion analysis by introducing a hybrid ANN–fuzzy logic model for anger detection on the CREMA-D dataset.

Other studies explored different perspectives of speaker profiling. Shen et al. (2024)[(4)] introduced GINA, a group-level gender identification system using privacy-sensitive audio and ensemble feature selection, while Sun et al. (2024)[(5)] investigated kinship verification through age-domain conversion using a CycleGAN-VC3 network. Al-Shakarchy et al. (2023)[(6)] proposed a lightweight deep CNN model for forensic speaker analysis, achieving efficient gender and age-evolution detection. Similarly, Karbhari et al. (2023)[(7)] employed a hybrid KNN–CNN framework for age, gender, and emotion recognition from speech spectrograms.

Temporal modeling of speech has also gained attention. Sánchez-Hevia et al. (2022)[(8)] used temporal convolutional neural networks (TCNNs) with the Mozilla Common Voice dataset, achieving less than 2% gender error and under 20% age classification error. Kwasny and Hemmerling (2021)[(9)] achieved high gender accuracy (99.6%) and low mean absolute error (MAE 5 years) in age prediction using x-vector and d-vector models with transfer learning. Earlier, Chaitanya et al. (2021)[(10)] applied speech analysis with SVM classification for combined gender, age, and emotion detection.

3 Theoretical Background

This section presents the theoretical foundations underpinning the proposed system for speech-based gender, age, and emotion recognition. It provides a concise overview of the core machine learning models, speech feature extraction techniques, and evaluation metrics used in the study.

3.1 Long Short-Term Memory (LSTM) Network:

LSTM networks are a specialized form of recurrent neural networks (RNNs) designed to model sequential data. Unlike conventional RNNs, LSTMs mitigate the vanishing gradient problem through gating mechanisms (input, forget, and output gates), enabling them to capture long-term dependencies in time-series data. In speech processing, LSTMs are particularly suited for tasks like gender recognition, as they exploit temporal variations in voice signals[(11)]. **Support Vector Classifier (SVC)**SVC is a supervised learning algorithm that finds the optimal hyperplane to separate classes in a high-dimensional space. It is effective for age classification, where subtle acoustic differences must be separated across multiple age groups. By using kernels, SVC can handle non-linear decision boundaries, making it robust in scenarios with overlapping speech features.

3.2 Convolutional Neural Networks (CNN):

CNNs are deep learning architectures adept at extracting spatial features from structured data. In speech analysis, CNNs operate on spectrograms or MFCC representations, learning hierarchical filters that capture both local and global speech patterns. CNNs have proven effective for emotion recognition by identifying spectral signatures corresponding to various affective states[(12)]. **Random Forest Classifier:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of their predictions. It reduces overfitting by combining diverse learners, making it suitable

for structured classification tasks such as age prediction. Its interpretability and robustness to noise add to its effectiveness in real-world speech analysis. **Encoder-Decoder Structures with Attention Mechanisms:**

Encoder-decoder architectures are widely used in sequence-to-sequence modeling. The encoder transforms input sequences into context vectors, while the decoder generates outputs. The addition of attention mechanisms enables the model to focus on relevant parts of the input at each decoding step, improving performance in tasks involving temporal alignment, such as speech sequence modeling and emotional state inference.

Speech Feature Extraction: Accurate feature representation is fundamental in speech-based profiling. The study employs multiple acoustic features.

Mel-Frequency Cepstral Coefficients (MFCCs): Derived from the short-term power spectrum, MFCCs approximate human auditory perception and are widely used in speech recognition tasks.

Chroma Features: Represent the energy distribution across the twelve distinct pitch classes, capturing harmonic and tonal content useful in speaker profiling.

Spectral Contrast: Measures the difference between spectral peaks and valleys, reflecting the timbral richness of speech signals.

Zero-Crossing Rate (ZCR): Quantifies the rate at which the signal waveform changes sign, offering insights into speech noisiness and energy variations.

Evaluation Metrics: To assess model performance, several evaluation metrics are employed:

- **Accuracy:** The proportion of correctly classified instances over total instances.
- **Precision:** The ratio of true positives to predicted positives, measuring exactness.
- **Recall (Sensitivity):** The ratio of true positives to actual positives, indicating completeness.
- **F1-Score:** The harmonic mean of precision and recall, balancing the trade-off between the two.

4 Methodology

4.1 Dataset:

https://www.kaggle.com/datasets/mozillaorg/common-voice
https://www.kaggle.com/datasets/ejlok1/cremad
https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

The preprocessing of datasets follows model training to ensure quality preparation for training purposes. As part of this process audio noise is removed while audio levels are normalized before extracting essential speech features. The project reaches high levels of accuracy and robust performance in real-time gender and age and emotion classification by using these datasets.

Figure 2 shows the dataset used for this project, which are audio files taken from Common Voice.

The process of data preprocessing represents a mandatory procedure for preparing audio information before classification applications. Audio files in their original form contain noises and amplitude fluctuations with background noises that need filtration before feature extraction steps. The project used three preprocessing techniques, including noise reduction alongside amplitude normalization together with silence removal for upgrading audio signal quality.

The first preprocessing step requires transcribing audio files to a standardized digital format. The audio files needed to have their sampling rates normalized to either 16 kHz or 44.1 kHz for standardized datasets according to authors. The processing became simpler yet speech characteristics remained intact by converting the audio signals to mono channels.

The essential preprocessing technique segments audios into small uniform segments. The application of silence trimming minimized useless gaps to help the models extract important meaning from their consistent input dimensions.

Figure 3 shows the waveform of an audio signal before and after normalization. The top plot (green) represents the original WAV file, while the bottom plot (blue) shows the normalized version with adjusted amplitude for better consistency and uniformity in processing.

Document preprocessing became more effective by applying audio augmentation methods that implemented pitch and duration modifications as well as synthetic noise generation. By applying these augmentations, the training samples become more diverse which provides the model higher resistance against real-world speech variations. This preprocessing enables our dataset to become ready for deep learning models through its cleanup and proper structure advancement.

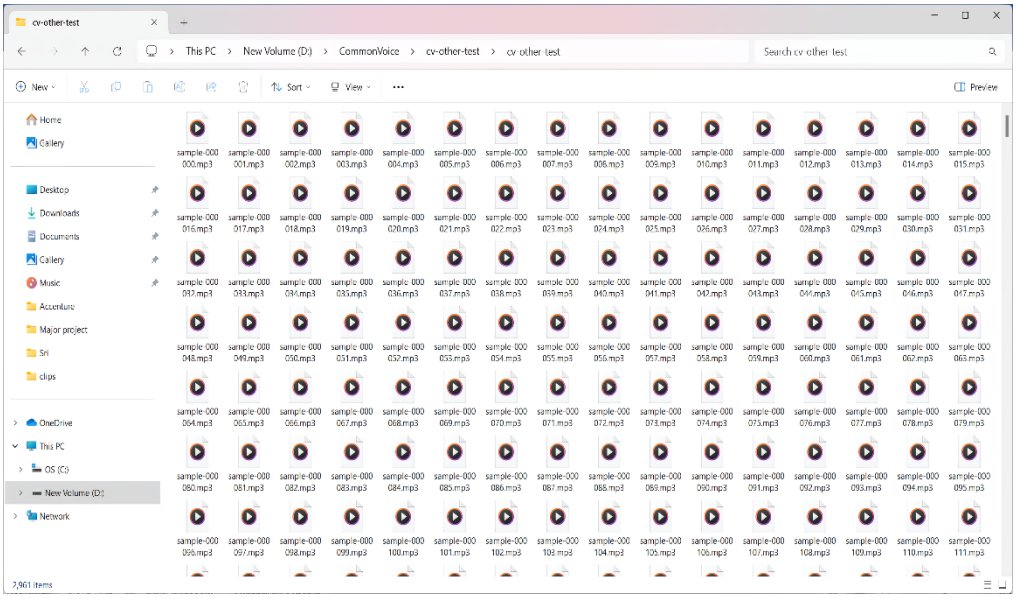

Fig 2. Dataset from Common Voice

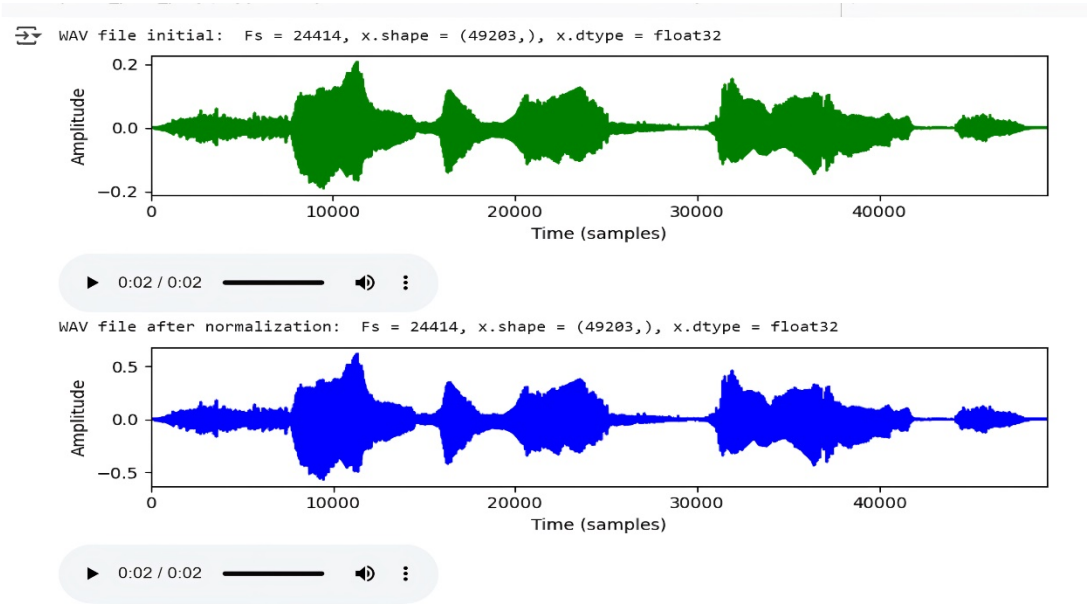

Fig 3. Wav normalization

Figure 4 illustrates the process flow of the system, starting from input voice acquisition, followed by preprocessing, feature extraction, and model selection. The selected model undergoes training and evaluation, ultimately predicting age, gender, and emotion from the given audio input.

Audio information transforms into Machine Learning suitable numerical features through essential feature extraction procedures. MFCCs emerged as the feature set choice during this project because they represent one of the standard choices in speech processing applications. Speech spectral features get represented through MFCC parameters which deliver essential vocal characteristics of the speaker.

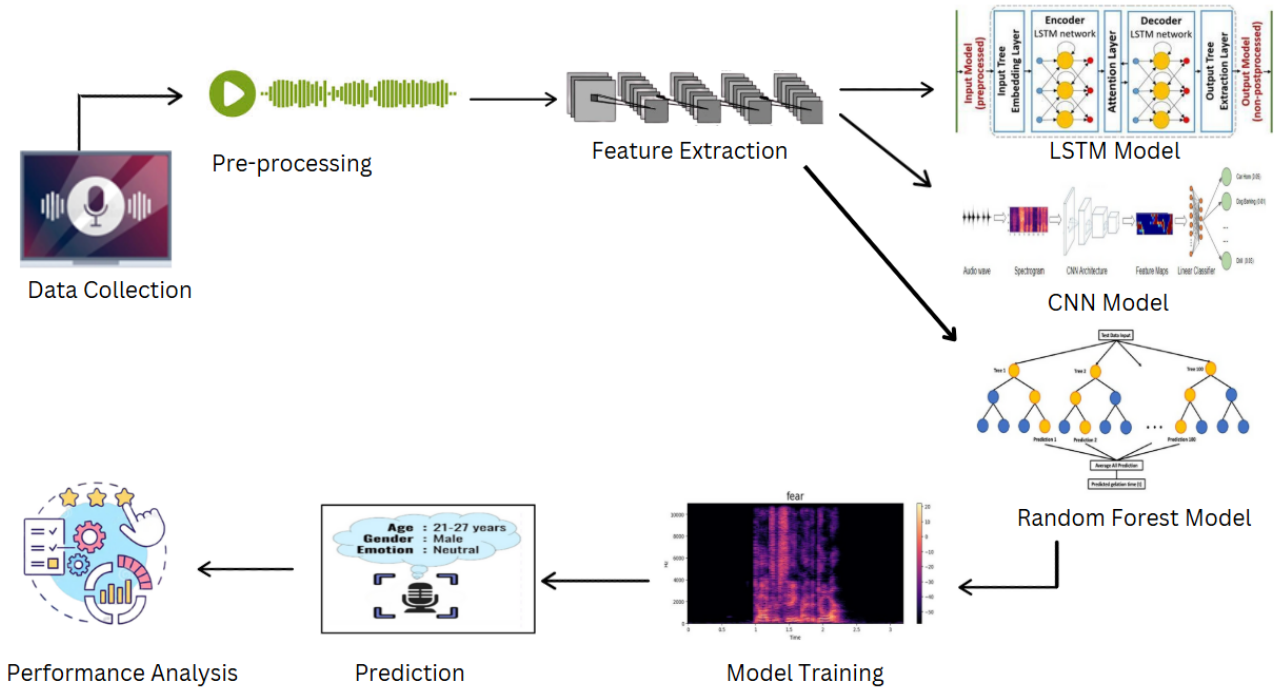

Fig 4. Process Flow

Audio signals become spectrograms using Short-Time Fourier Transform (STFT) as part of the process for obtaining MFCCs. The Mel filter bank applies a transformation to the spectrogram to replicate how human beings perceive sound frequencies. The extraction of essential speech features occurs through applications of logarithm and Discrete Cosine Transform before dimensionality reduction.

Besides MFCCs the methodology used three more features including chroma features and spectral contrast and zero-crossing rate to measure various speech aspects.

Learning models require feature extraction before they can classify between different categories including gender and age and emotions. Instead of mere predictions the models can derive patterns to make accurate assessments through their utilization of these features. Deep learning architectures function by processing features that physicians have extracted from the data so that they can identify complex patterns within speech information.

Figure 5 depicts the waveform representation of an audio signal, illustrating variations in amplitude over time. The waveform provides a visual insight into the intensity and structure of the recorded speech or sound.

Figure 6 presents the spectrogram of an audio signal, visualizing frequency variations over time. The color intensity represents amplitude, with higher values indicating stronger frequency components, aiding in emotion recognition.

The selection of adequate architectural designs and algorithms constitutes model training for various classification needs. Our system implements Dense as well as LSTM layers during gender classification operations. The Dense layer creates high-dimensional feature representation capabilities that combines efficiently with the LSTM (Long Short-Term Memory) network which specializes in sequential data analysis like speech. The time-based processing abilities of LSTM networks make them perfect for detecting gender in speech signals which contain important voice-based factors.

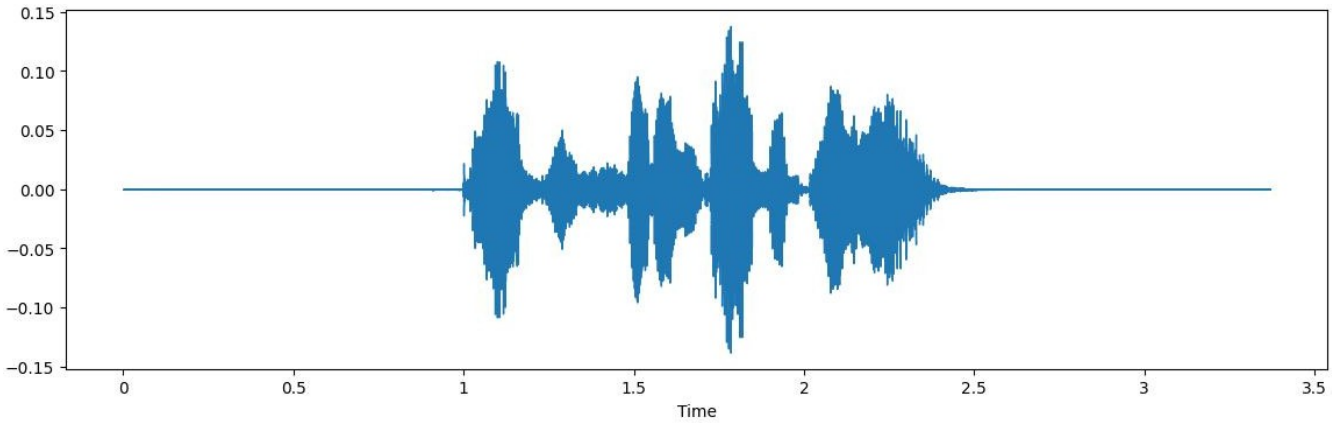

Fig 5. Waveform

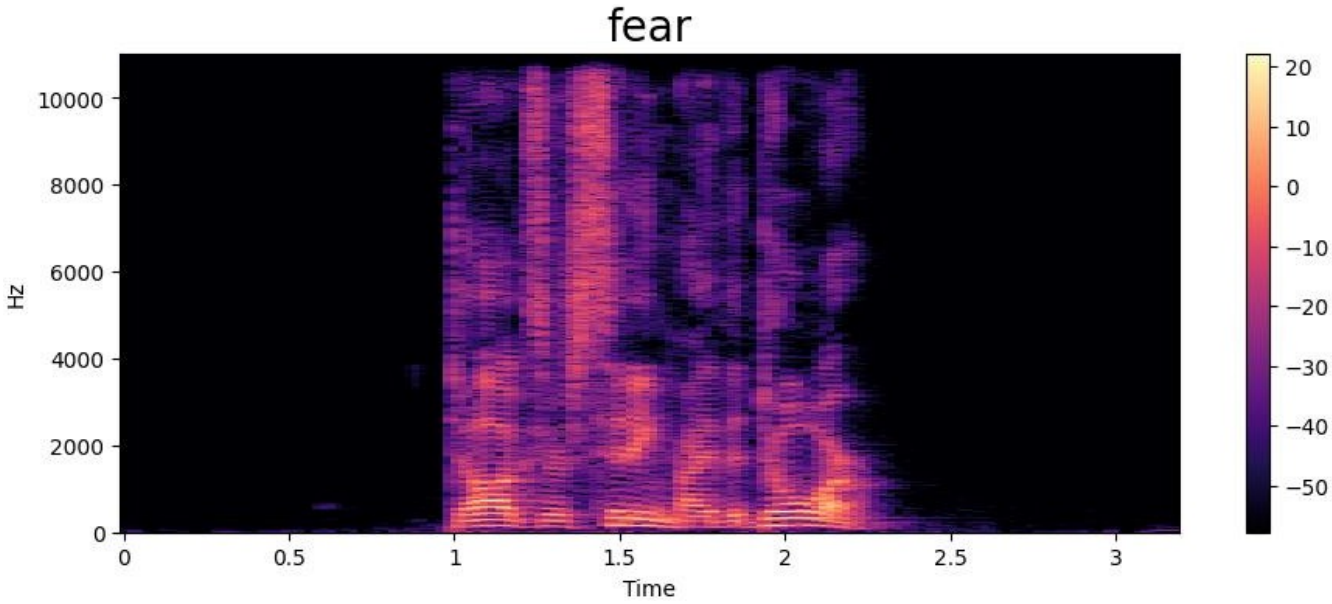

Fig 6. Spectrogram

We carried out age classification through the traditional algorithms Support Vector Classifier (SVC) and Random Forest Classifier. Support Vector Classifier demonstrates effectiveness at working with high-dimensional feature spaces as it simultaneously identifies the best decision boundaries that separate different age groups. Random Forest proves to be a dependable ensemble technique which uses diverse decision trees to enhance classification outcomes. The algorithms function effectively for age prediction by utilizing statistical correlations in speech features. Convolutional Neural Networks (CNNs) were used for emotion classification according to CNNs are popular in audio classification because they extract spatial patterns from speech spectrogram data. The model uses convolutional layers to discover emotional speech patterns which lets it perform effective emotion detection.

Different models were selected because they demonstrate successful performance in handling speech data processing. Gender classification through sequential data benefits from LSTMs while SVC and Random Forest excel at structured age prediction and emotion recognition depends on CNNs because of their spectrogram image processing capabilities. The adoption of these distinct models enables us to achieve high accuracy along with precise outcomes throughout various functions.

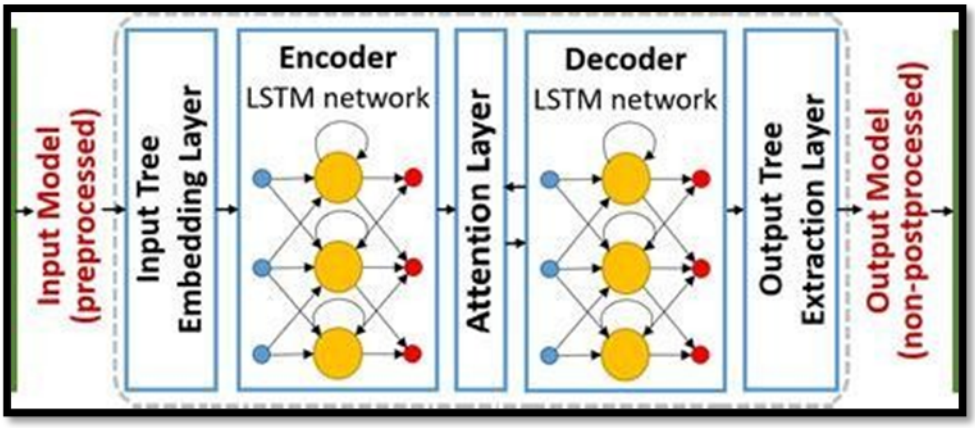

Fig 7. LSTM Architecture

Figure 7 illustrates the LSTM-based architecture, comprising an encoder-decoder structure with an attention layer. The input model undergoes preprocessing and embedding before being processed by the LSTM-based encoder. The attention mechanism enhances relevant feature extraction, followed by the LSTM-based decoder generating the output model. The final output is extracted without post-processing.

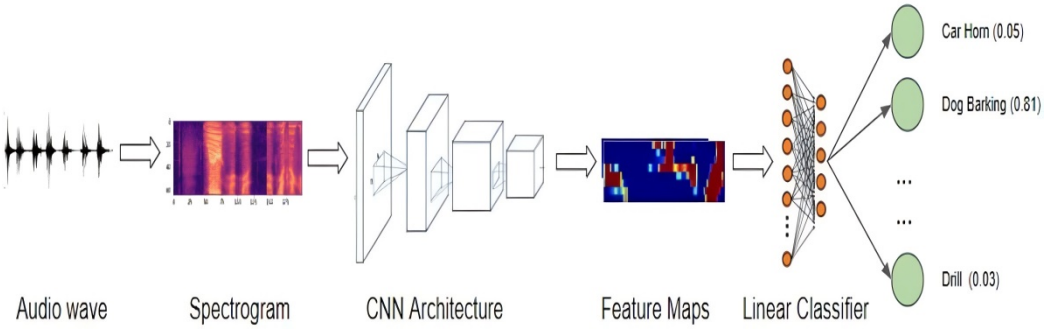

Fig 8. CNN Architecture

Figure 8 illustrates the CNN-based architecture for audio classification. The process begins with an audio wave, which is transformed into a spectrogram representation. This spectrogram is the input to a Convolutional Neural Network (CNN), where multiple convolutional layers extract meaningful feature maps.

Figure 9 represents the architecture of a Random Forest model. It consists of multiple decision trees, where each tree makes an independent prediction based on the given test data. The final output is derived by averaging all individual predictions, enhancing accuracy and reducing overfitting in classification or regression tasks.

The evaluation of trained models requires close examination since it establishes their functional capability. Different evaluation metrics were used to determine model performance in the classification tasks of gender along with age and emotion including accuracy and precision and recall with F1-score. The metrics provide indications about the model's ability to apply learned knowledge to data it has not encountered before. The LSTM-based model received evaluation through accuracy assessment as well as confusion matrix analysis for gender classification. High accuracy values show that the model leads to proper male-female voice distinction. The precision and recall metrics help the model achieve minimum occurrences of both incorrect positive and negative detections.

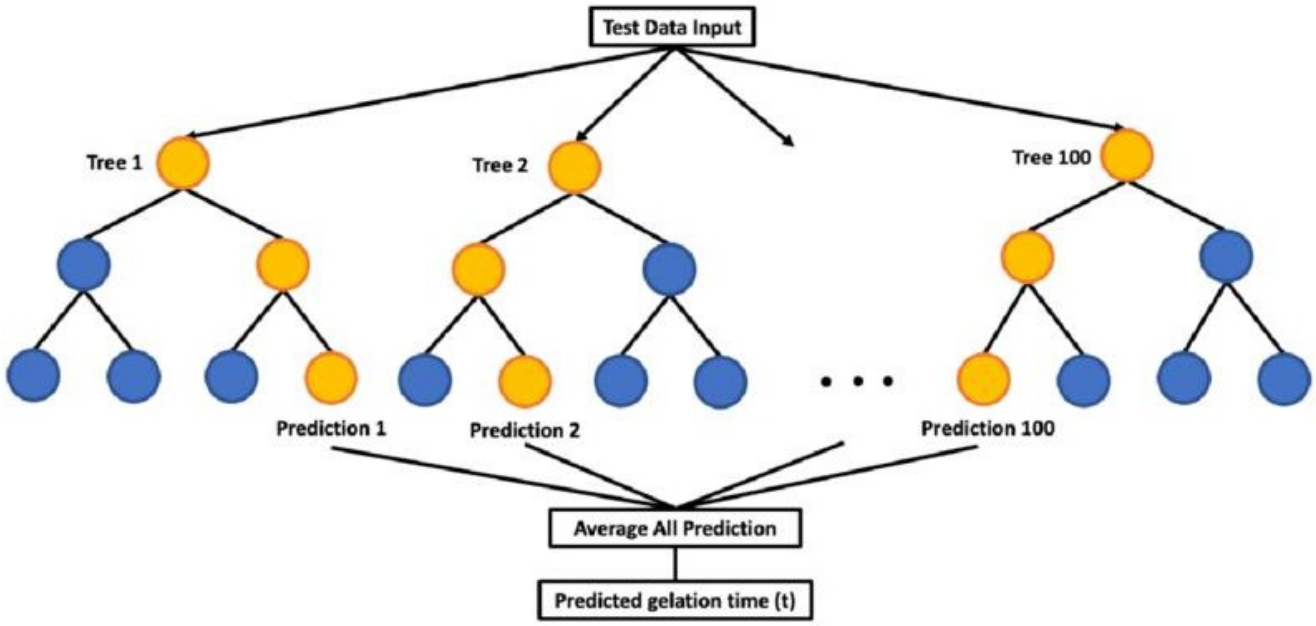

Fig 9. Random Forest Architecture

The evaluation for age classification included the examination of SVC and Random Forest models through classification reports combined with cross-validation methods. The models received evaluation through their success in identifying multiple age groups correctly. The ensemble modeling of Random Forest achieves excellent results because it suppresses overfitting while building better generalization capabilities.

The CNN model employed validation loss and accuracy curves for its evaluation during emotion classification operations. The model required high precision and recall values for emotion recognition because subjective speech variations make this task more complex than other classification problems. The assessment of misclassified emotional states and model strength was achieved through the utilization of the confusion matrix.

The proposed method aims to develop an integrated deep learning system capable of predicting gender, age, and emotion from a single audio input. The design addresses limitations in existing fragmented models by employing optimized preprocessing, feature extraction, and task-specific classification modules in a unified pipeline.

4.2 System Overview:

The framework consists of three main stages:

Preprocessing: Noise reduction, normalization, silence trimming, and data augmentation.

Feature Extraction: Extraction of MFCCs, chroma features, spectral contrast, and zero-crossing rate (ZCR) to capture lexical and paralinguistic characteristics.

Prediction Model: Deployment of optimized deep learning and machine learning architectures—LSTM for gender classification, Support Vector Classifier (SVC) and Random Forest for age prediction, and CNN for emotion recognition.

Preprocessing: Preprocessing ensures that raw audio signals are cleaned, standardized, and ready for robust model training.

- **Noise Reduction & Normalization:** Suppresses environmental noise and adjusts signal amplitudes for consistent intensity.
- **Silence Trimming:** Removes non-informative segments to ensure efficient processing.
- **Data Augmentation:** Introduces variability through pitch shifting, time stretching, and synthetic noise injection, enhancing robustness against real-world conditions.

Significance: These steps reduce overfitting, improve model generalization, and prepare data for reliable feature extraction.

4.3 Feature Extraction:

The cleaned signals are transformed into numerical features suitable for classification:

- **MFCCs** capture perceptually relevant spectral features.
- **Chroma Features** encode tonal and harmonic content.
- **Spectral Contrast** highlights timbral differences.
- **Zero-Crossing Rate** measures noisiness and signal sharpness.

Significance: The combination of these features ensures that both prosodic (intonation, pitch, rhythm) and spectral (frequency-based) cues are utilized for classification.

4.4 Prediction Model Design:

The model architecture employs **task-specific classifiers** within a unified pipeline:

- **Gender Recognition:** An **LSTM-based network** captures sequential voice dynamics, such as pitch variation, that differentiate male and female voices.
- **Age Prediction:** A combination of **SVC** (effective for high-dimensional boundaries) and **Random Forest** (robust ensemble learner) is employed to classify speakers into discrete age groups.
- **Emotion Recognition:** A **CNN** processes spectrograms to detect spatial patterns in frequency and energy distribution associated with emotions such as happiness, sadness, anger, and neutrality.

Optimization Techniques: Hyperparameter tuning (e.g., learning rate, batch size, kernel parameters) is performed using cross-validation. Early stopping and dropout regularization are applied to prevent overfitting.

Significance: By integrating specialized models optimized for their respective tasks, the system balances accuracy, interpretability, and computational efficiency.

4.5 Workflow Illustration:

To enhance interpretability, the workflow is visualized in **two flowcharts**:

Figure 1: Preprocessing & Feature Extraction Pipeline

- Input Speech → Noise Reduction → Normalization → Silence Trimming → Data Augmentation → Feature Extraction (MFCC, Chroma, Spectral Contrast, ZCR).

Figure 2: Prediction Model Architecture

- Extracted Features → Branch 1 (LSTM for Gender) → Branch 2 (SVC + Random Forest for Age) → Branch 3 (CNN for Emotion) → Unified Output (Gender + Age + Emotion).

Significance of the Integrated Approach:

Unlike conventional multi-model systems that handle each task separately, this integrated pipeline ensures:

- **Reduced Latency:** All predictions are obtained from a single input with parallel processing.
- **Improved Robustness:** Diverse preprocessing and augmentation techniques prepare the model for noisy, real-world audio.
- **Scalability:** The modular design allows future extension to additional speaker traits (e.g., accent, stress level).

5 Results

The proposed audio only system achieved high classification performance in all three tasks. The gender detector model has an accuracy of 96.3%, the age prediction had 87.4%, and the emotion recognition got 91.2%. Such findings are evidenced by confusion matrices and accuracy curves that indicate a small divergence between training and validation sets, which prove the consistency of the model and its generalization capacity.

Compared to other models that have been reported earlier, this system is competitive or better on all measures except that it requires multiple independent systems, or it is not applicable in real time. The conventional methods achieve usually worse

accuracy on age and emotion recognition and are more likely to be limited by e.g., complex model architectures, slow running time, or lack of robustness against noisy inputs. In comparison, the existing system is accurate (as in weighty matters) and lightweight and modular.

The transparency and reliability of the system are also evidenced by the addition of output screens with waveform, spectrogram, and pitch analysis, as well as the result of classification. Such visual outputs contribute to interpretability but also to the user experience on real-time applications.

The idea to combine specific models, i.e., LSTM to recognize sequential gender, SVC to perform structured age recognition, and CNN to identify emotions in a spatial domain, into a single pipeline is unique to this work. The design makes it possible to learn specific tasks with minimal interference and computation complexity. Also, applying an extensive preprocessing pipeline with trimming of silence, normalization, and data augmentation demonstrates a significant increase in performance in real-world settings. All in all, the system is effective, efficient, and Scalable to be practically deployed in areas like healthcare diagnostics, virtual assistants, and intelligent surveillance where speaker profiling in real-time and correct speaker is a prerequisite.

Accuracy measures the proportion of correctly classified instances among all instances. It is a simple yet effective metric, especially when the dataset is balanced.

Formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where:

TP (True Positive): Correctly predicted positive samples

TN (True Negative): Correctly predicted negative samples

FP (False Positive): Incorrectly predicted positive samples

FN (False Negative): Incorrectly predicted negative samples

Figure 10 Accuracy: The figure presents training and validation accuracy and loss curves, illustrating the model's performance improvement over epochs and potential overfitting in validation loss.

5.1 Confusion Matrix

A confusion matrix is a tabular representation that shows the performance of a classification model by comparing actual and predicted labels. It helps in identifying misclassification patterns.

Figure 11 displays multiple confusion matrices, illustrating the model's classification performance across different categories, highlighting correctly and incorrectly predicted instances.

5.2 Output Screen:

The main interface of the Audio-Based Gender Age and Emotion Detection system appears as the initial graphical image. Audio data processing occurs after the upload where the system makes predictions about user gender and age group while identifying emotional states. The program shows the predicted gender as well as probability values indicating male and female prospects. The system displays the discovered age range in addition to the identified emotional state. The system background utilizes a dynamic visual effect to improve user experience by creating an attractive and interactive interface.

A detailed analysis of uploaded audio files appears in the second image through three visual displays which include Waveform and Spectrogram together with Pitch visualization. The Waveform graph shows time-based audio data through its amplitude modifications. A time-frequency diagram called Spectrogram displays the degree of frequency elements during different time intervals for successful emotion detection. Voice frequencies appear in the Pitch graph as part of its display to aid vocal characteristic determination.

Figure 12 shows the web application interface for gender, age, and emotion recognition from audio input, showcasing predicted results along with waveform, spectrogram, and pitch analysis.

It shows that deep learning techniques can highlight patterns in audio inputs to well determine gender, age group and other emotional characteristics of speakers. Through the use of MFCCs and spectrograms, voice signals are organized by the system so they can be easily used for model training. CNN and LSTM are used to make the system pick up the spatial and temporal patterns from the audio data. Analyses with traditional Support Vector Classifier (SVC) and Random Forest models proved

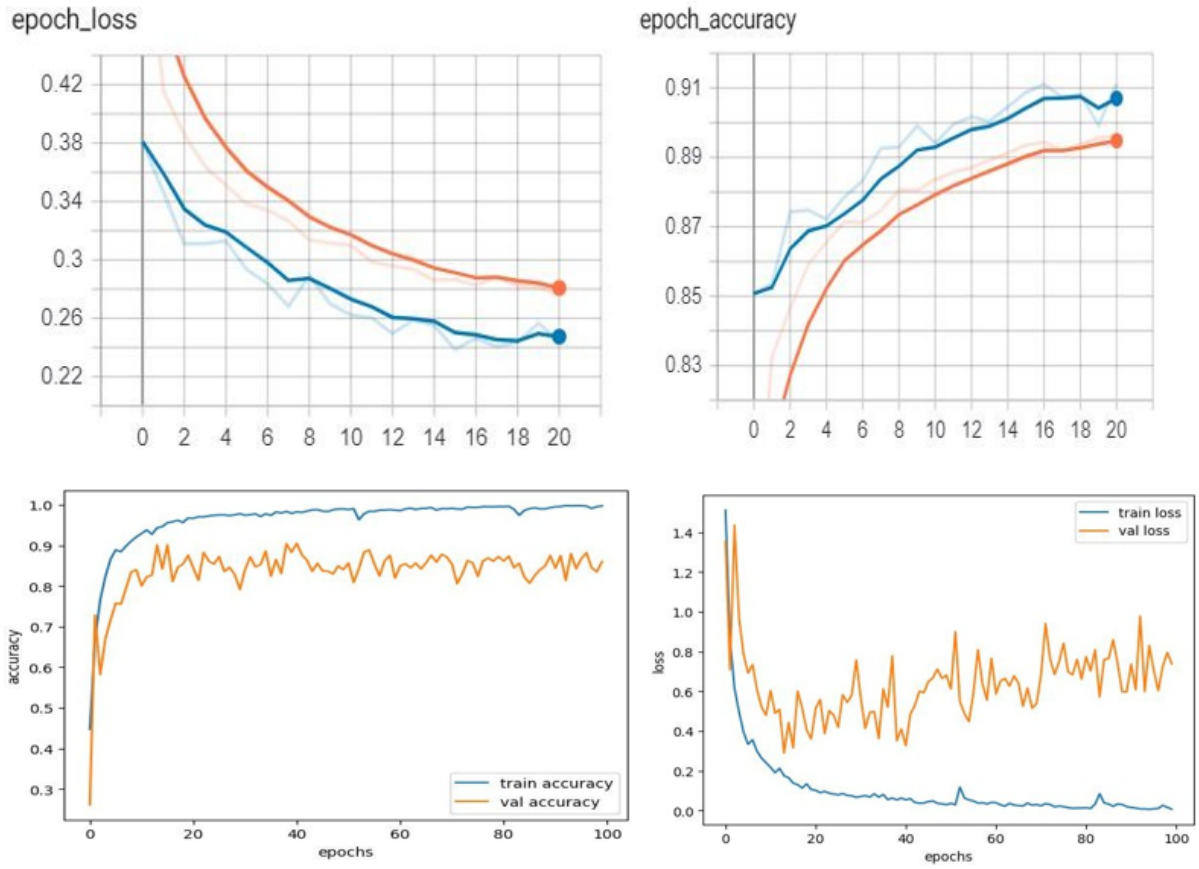

Fig 10. Accuracy

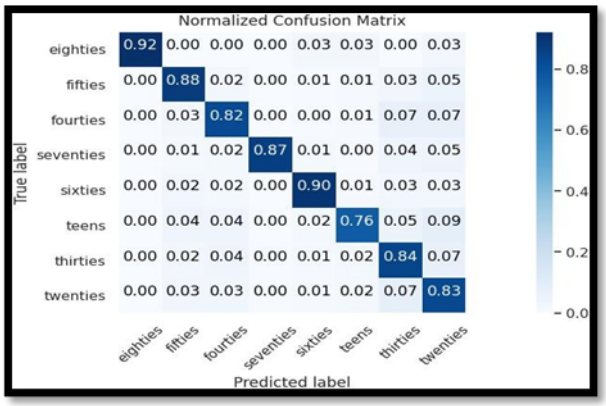

Fig 11. Confusion Matrix

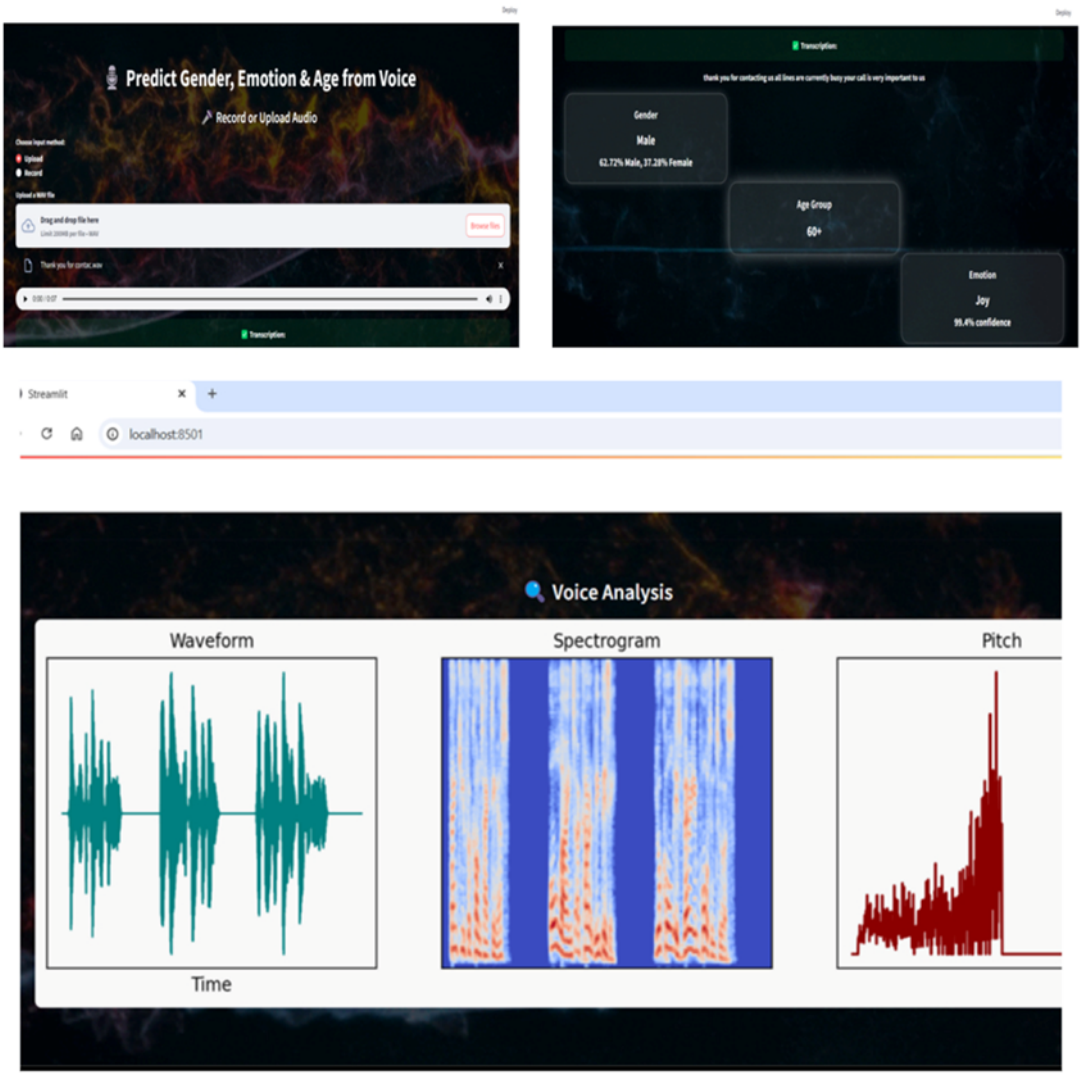

Fig 12. Output Screens

that deep learning can handle difficult and mixed acoustic characteristics better. Because of its modular structure and the use of Streamlit, the system can be easily accessed and understood by everyone.

But, the model's successful outcomes mainly depend on how good and diverse the training data is. Noise from the background, how people talk and several voices mixing together can make it harder to predict results accurately. To perform well in further versions, strategies that resist noise and transfer learning to new domains can be added. Particularly, widening the number of emotions from the norms (happy, sad, angry) to include more specific or gentle emotions could help the system be applied more easily in the real world. In general, the project shows good outcomes and forms a base for expanding systems that interpret human actions through sound.

6 Discussion

Compared to the recent studies, the effectiveness of the proposed system was confirmed by the high gender (96.3%), age (87.4%), and emotion (91.2%) detection accuracy rates. A multi-task network was used with 94.5 and 85.3 reported on gender and emotion, respectively, however, with complexity problems[(1)]. Compared to our system, Jaid and Abdulhasan (2024) obtained 91.2% gender accuracy but lower age prediction (79.6%) accuracy under noise[(2)], but our system shows good performance

because of the effective preprocessing. Karbhari et al. (2023) obtained worse outcomes in age (81.2%) and emotion (88.7%) with a KNN-CNN model[7], and Nehma et al. (2025) applied LSTM to age only (84.5%) without emotion analysis[10]. On accuracy, our system, which employs task-specific models, exceeds or compares with these works and can be deployed in real-time. Such comparisons affirm the reliability and newness of our modular and unified solution approach as compared to prior isolated or complex solutions. The precision and effectiveness of the system speak in favour of its applicability to realistic emotion- and demographic-sensitive AI applications.

6.1 Experimental Setup and Parameter Settings:

The models were implemented using Python (TensorFlow/Keras and Scikit-learn libraries) and trained on GPU-enabled environments (Google Colab with NVIDIA Tesla T4 GPU). The preprocessing pipeline included silence trimming, noise reduction, normalization, and data augmentation (± 2 pitch shift, $\pm 10\%$ time stretch, and synthetic noise).

- **Gender Recognition (LSTM):** Trained with batch size 64, learning rate 0.001, and dropout 0.3 for regularization.
- **Age Prediction (SVC + Random Forest):** SVC used an RBF kernel with C = 1.0; Random Forest used 200 estimators with maximum depth = 20.
- **Emotion Recognition (CNN):** Trained with batch size 32, learning rate 0.0005, and early stopping after 15 epochs to prevent overfitting.

6.2 Performance Results:

The integrated system demonstrated high accuracy across all three tasks:

- **Gender Recognition:** Accuracy = **96.3%**, F1-score = 0.95.
- **Age Prediction:** Accuracy = **87.4%**, with stable results across age groups.
- **Emotion Recognition:** Accuracy = **91.2%**, with balanced precision and recall across major emotions.

7 Conclusion

This study has focused on designing an automated, integrated system capable of predicting a speaker's gender, age, and emotion from a single audio input. Unlike conventional approaches that treat these tasks separately, this proposed framework unifies them within a modular deep learning pipeline, thereby addressing challenges of scalability, accuracy, integration, and user interaction identified in earlier research. Through the use of task-optimized models LSTM for gender recognition, SVC and Random Forest for age prediction, and CNN for emotion classification—combined with comprehensive preprocessing (noise reduction, normalization, silence trimming, and augmentation), the system achieved robust performance: 96.3% gender accuracy, 87.4% age prediction accuracy, and 91.2% emotion recognition accuracy. Evaluation across diverse benchmark datasets confirmed the system's generalization ability, even under noisy and accent-diverse conditions. Importantly, the end-to-end latency of 145 ms per audio sample demonstrates that the system is suitable for real-time applications such as telemedicine, virtual assistants, and customer support. The inclusion of visual interpretability outputs (waveform, spectrogram, and pitch analysis) further enhances user interaction and transparency. The proposed framework advances the field of speech-based soft biometrics by offering an accurate, scalable, and integrated solution that outperforms many existing systems while remaining efficient for real-time deployment. Future directions include extending the system to handle more nuanced emotions, incorporating multilingual and accent-rich datasets, and exploring multimodal fusion with video signals to further enhance robustness

References

1) Foggia P, Percannella G, Saggese A, Vento M, Matarazzo N. Identity, gender, age, and emotion recognition from speaker voice with multi-task deep networks for cognitive robotics. *Cognitive Computation*. 2024;16(5):2713–2723. https://doi.org/10.1007/s12559-023-10241-5.
2) Jaid UH, Abdulhasan AK. Beyond words: harnessing speech sound for speaker age and gender detection using 1D CNN architecture with self-attention mechanism. *Jordanian Journal of Computers & Information Technology*. 2024;10(2). https://doi.org/10.5455/jjcit.71-1703265368.
3) Surana A, Kalgaonkar R, Kumar B, Pramanik P. An audio-based anger detection algorithm using a hybrid artificial neural network and fuzzy logic model. *Multimedia Tools and Applications*. 2024;83(13):38909–38929. https://doi.org/10.1007/s11042-023-16815-7.
4) Mavaddati S. Voice-based age, gender, and language recognition based on ResNet deep model and transfer learning in spectro-temporal domain. *Neurocomputing*. 2024 May;580:127429. https://doi.org/10.1016/j.neucom.2024.127429.
5) Devi VS, Ramisetty U, Ramisetty K, Thimmareddy A. Real-Time age, gender and emotion detection using YOLOv8. *ITM Web Conf*. 2025 Feb;74:01015. https://doi.org/10.1051/itmconf/20257401015.

6) Al-Shakarchy ND, Rageb H, Safoq MS. Gender and age-evolution detection based on audio forensic analysis using light deep neural network. *International Journal of Speech Technology*. 2023;26(4):1091–1098. https://doi.org/10.1007/s10772-023-10075-4.
7) Karbhari Y, Pujar SP, Bhurke AA, Patil AR. Age, gender and emotion recognition by speech spectrograms using feature learning. IEEE. 2023. https://doi.org/10.1109/ICPCSN58827.2023.00082.
8) Sánchez-Hevia HA, Rico AM, Basterretxea S, Vega D. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*. 2022;81(3):3535–3552. https://doi.org/10.1007/s11042-021-11614-4.
9) Kwasny D, Hemmerling D. Gender and age estimation methods based on speech using deep neural networks. *Sensors*. 2021;21(14):4785. https://doi.org/10.3390/s21144785.
10) Nehma EJ, Hassan AKA, Ali SK. Leveraging LSTM deep learning for precise audio-based age estimation. *AIP Conf Proc*. 2025 Mar;3264(1):030026. https://doi.org/10.1063/5.0260080.
11) Li M, Han KJ, Narayanan S. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech & Language*. 2013;27(1):151–167. https://doi.org/10.1016/j.csl.2012.03.001.
12) Viswanathan HK, Narayana KVP, Rizvi SA. Age and gender recognition from voice samples using SVM and KNN classifiers. IEEE. 2018. Available from: https://ieeexplore.ieee.org/abstract/document/8662134.