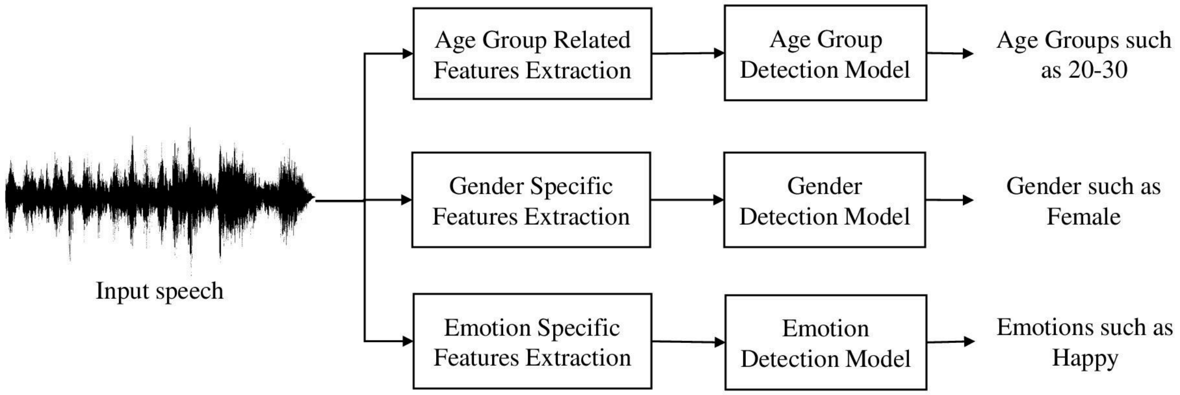
Age, Gender and Emotion Detection from Speech Using AI

Introduction and Background:

Speech recognition technologies have garnered substantial research attention owing to their manifold applications in domains spanning human-computer interaction, virtual assistants, and affective computing systems. The present work examines the challenging undertaking of age-group, gender, and emotion identification from vocal signals, contending with complications introduced by intricate variabilities and divergences inherent to human speech modalities across various demographic and affective dimensions. Detecting age group, gender, and emotion using individual models, which is similar to three parallel processes, presents a number of issues that may impede the efficiency and efficacy of voice-based identification systems. These difficulties include higher latency caused by the simultaneous execution of three independent models, greater memory requirements caused by executing multiple models in parallel, and the possibility of inaccuracies in one model's predictions affecting the outcomes of the others. The development of advanced learning techniques, particularly Convolutional Neural Networks (CNNs) that use 1D convolutions have changed the field of audio signal processing. In the pursuit of advancing the frontier of voice-based recognition systems, our research offers a thorough examination employing a spectrum of models for age-group, gender, and emotion identification. Additionally, integrated models are explored to unravel the intricate links between gender and age-group, as well as gender and emotion. The overarching aim is to contribute to the evolution of contemporary voice recognition technology by furnishing profound insights into acoustic cues indicative of age-group, gender, and emotional states. Datasets utilized in this investigation hail from diverse sources, including the Mozilla Common Voice Dataset (Mozilla). These datasets encompass a rich variety of speech samples, enabling robust model training and evaluation across a wide spectrum of demographic and emotional scenarios. The major challenge faced in the “Combined Model” approach is the unavailability of feature datasets which have all the 3 output parameters namely :- age-group - Group, Gender, Emotion.

Voice-based recognition systems have received a lot of attention, and multiple research projects have looked into different aspects of age group, gender, and emotion detection using voice data. We present a selection of significant works that provide insights into the world of voice-based recognition, notably in the context of age-group, gender, and emotion detection. These publications provide significant perspectives on various approaches, obstacles, and developments, establishing the framework for our complete research aimed at recognising age-group, gender, and emotion from voice data: 1. The work of S. R. Zaman et al. adopts a distinctive perspective in the domain, employing audio speech as a singular source for concurrent gender, age-group, and mood recognition. A range of models, including CatBoost, Random Forest, and XGBoost, undergo testing with 20 statistical characteristics. Notably, CatBoost attains a remarkable 96.4% accuracy in gender prediction, Random Forest excels with 70.4% in age-group prediction, and XGBoost leads with 66.1% in emotion prediction. The scrutiny of these key elements furnishes valuable insights, charting a course for future research in voice-based recognition. 2. In the research conducted by Héctor A. Sánchez-Hevia et al. , Deep Neural Networks undergo evaluation for the joint prediction of age-group and gender from speech—a crucial aspect for Interactive Voice Response (IVR) systems in contact centers. Leveraging Mozilla's Common Voice dataset, the findings reveal resilient gender classification across networks, with larger sizes contributing to enhancement. A combination of convolutional and temporal neural networks emerges as the optimal configuration for age-group group classification, showcasing potential for IVR systems with minimal gender identification error (below 2%) and age-group group classification error (below 20%) in the most effective systems. 3. In the research conducted by Poonam Rani et al., the research introduces a system designed for discerning an individual's emotional state through audio signal registrations, with potential applications in speech analytics and personalized human-machine interactions. The investigation encompasses two datasets, each comprising approximately 3000 speech samples for gender analysis and 1000 samples for emotion evaluation. 4. In the research conducted by Lee et al., the focus is on multitask learning for concurrent speaker age-group and gender classification, showcasing the efficacy of shared representations. 5. In the work presented by Gómez et al., a novel approach is taken to address deficiencies in vocal pathology detection systems by introducing an age-group detector trained with both normal and disordered voices. Concentrating on adults and the elderly, the research leverage-groups Mel frequency cepstral coefficients and Gaussian mixture models sourced from the Saarbruecken database. Notably, the research attains an impressive accuracy of 96.57%, showcasing the potential for developing autonomous age-group dependent speech pathology identification systems. 6. In the research conducted by Prasanta et al., a Tensor-based strategy is introduced for the detection of speaker gender in speech-based communication, a pivotal aspect in enhancing voice recognition systems. The proposed technique employs a GMMbased classifier tailored for low-resource language-groups. Through experiments conducted on the TIMIT and SHRUTI datasets, the research attains an average-group gender detection accuracy of 91%. The analysis of these results underscores the effectiveness of the Tensor-based approach in the precise detection of speaker gender. 7. In the research conducted by Zheng et al., the research takes aim at mitigating challenges in emotion recognition arising from speaker variability and limited training samples. A novel solution is proposed in the form of a context-dependent domain adversarial neural network (DANN) designed for multimodal emotion recognition. Emphasizing the importance of contextual information and multimodal features, the method strives to predict emotion labels while concurrently learning a common representation that diminishes disparities related to speaker identity. To address the limitations of low-resource samples, the strategy incorporates unlabeled data. Experimental results on the IEMOCAP dataset showcase an absolute improvement of 3.48% over state-of-the-art techniques, affirming the efficacy of the proposed approach.



Age Group Related Features Extraction
Age Group Detection Model
Age Groups such as 20-30
Gender Specific Features Extraction
Gender Detection Model
Gender such as Female
Input speech
Emotion Specific Features Extraction
Emotion Detection Model
Emotions such as Happy

Fig 1. Conventional Approach

Problem Analysis and Solution Development:

The issue of identifying age group, gender, and emotion from audio data presents considerable hurdles due to the complexity of human speech modalities and the necessity for accurate and rapid recognition algorithms. Human speech varies intricately across demographic and emotional aspects, making it impossible to reliably identify age groups, genders, and emotions merely based on vocal signals. These variables include a wide range of parameters such as accent, pronunciation, intonation, and speech patterns, all of which contribute to the nuances of speech recognition tasks. Addressing these variables necessitates strong models capable of successfully capturing and interpreting these complexities. Furthermore, the availability of labeled datasets including all three output factors - age-group, gender, and emotion - is restricted, making it difficult to train comprehensive recognition algorithms. The scarcity of high-quality, diversified training data makes it difficult to construct models that are generalizable across different demographic and emotional contexts. Ensuring the quality and diversity of training data is critical for improving the robustness and generalizability of voice recognition systems. Previous techniques that relied on different models for each task have increased latency due to the simultaneous execution of independent models, which limits real-time applications of voice-based recognition systems. Adding many jobs to a single model raises concerns about model complexity and computational expense. Efficient systems capable of performing multi-label classification jobs are required to address these issues and enable real-time audio data processing. The dynamic nature of human speech, as well as the context-dependent variances that exist within different demographic and cultural groups, make identifying age groups, genders, and emotions from audio data much more difficult. Speech patterns, language usage, and linguistic nuances vary between age groups, genders, and emotional states, adding complexity to the recognition process. These variances may manifest differently in different languages, dialects, and socio-cultural situations, necessitating the creation of models that can adapt and generalize across diverse linguistic and cultural landscapes. Furthermore, the existence of noise, distortions, and environmental factors that might affect the quality of the audio signal makes it more difficult to reliably identify age groups, genders, and emotions from audio data. Background noise, microphone artifacts, and acoustic reverberations can all interfere with the extraction of key characteristics, reducing the effectiveness of recognition models. Robust feature extraction techniques and noise-tolerant modeling approaches are required to reduce the impact of these environmental factors and improve the reliability of recognition systems. Potential biases in the training data should also be taken into account as they may cause differences in the model's performance among various demographic groups. Unfair representation of particular demographic groups, cultural prejudices in the annotation labeling process, or innate flaws in the data collection procedure can all contribute to biases in the training data. In order to mitigate these biases and guarantee that recognition systems operate fairly and equally across a range of demographics, rigorous validation processes, thorough training data curation, and the implementation of bias mitigation strategies are necessary.

Furthermore, the use of voice recognition technologies to identify emotions, age groups, and genders brings up significant privacy and security issues with relation to the gathering, storing, and using of private voice recordings. Building confidence and trust in speech recognition systems requires establishing strong security measures, protecting user privacy, and guaranteeing data confidentiality. Adherence to ethical norms and data protection rules is crucial in safeguarding user rights and reducing the likelihood of illegal access or exploitation of confidential voice data.

The problem of identifying age group, gender, and emotion from audio data is made more difficult by the absence of established metrics and evaluation methodologies for evaluating recognition system performance, on top of the previously described difficulties. There may be disparities between objective performance measures and subjective human impressions as a result of existing evaluation metrics' inability to adequately reflect the subtleties and complexity of age-group, gender, and emotion prediction tasks. To assure the efficacy and dependability of speech recognition systems across a wide range of applications and use cases, thorough assessment frameworks that include several dimensions of recognition accuracy, fairness, robustness, and user satisfaction are important. Moreover, temporal dependencies and contextually sensitive information are introduced by the dynamic nature of human communication, and recognition systems need to collect and characterize these effectively. Due to their inherent sequential structure, speech signals provide important information regarding age group, gender, and mood through temporal patterns and dependencies. Advanced sequence modeling techniques, such as transformers or recurrent neural networks (RNNs), that can capture long-range dependencies and contextual information across audio sequences are needed to describe these temporal dynamics. Contextual data from surrounding utterances, speaker interactions, and environmental signals can be incorporated into recognition models to improve their robustness and discriminative power, resulting in predictions that are more accurate and contextually aware.

To sum up, the process of identifying age groups, genders, and emotions from audio data is a complicated and multidimensional issue that calls for resolving a wide range of ethical, cultural, technological, and privacy concerns. We can create trustworthy, equitable, and transparent voice recognition systems that advance affective computing and human-computer interaction by utilizing sophisticated modeling techniques, strong feature engineering, thorough evaluation frameworks, and ethical data practices. In addition to a dedication to openness, justice, and user privacy, developing trust and confidence in speech recognition systems calls for continuous study and cooperation to take advantage of new opportunities and difficulties in this quickly developing sector.

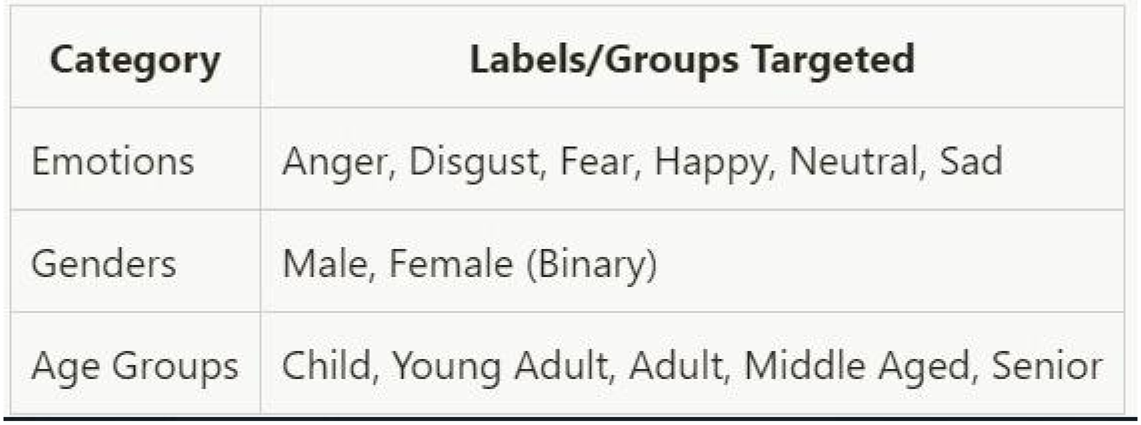
Proposed Solution:

The proposed solution, illustrated in the diagram [1], adopts a deep learning pipeline for multi-output emotion recognition. This method involves individual Convolutional Neural Networks (CNNs) processing audio information extracted from speech to predict emotion, gender, and age-group. While the "Emotion detection Model" improves emotion detection by adding age-group and gender estimates, the "Single Multi-output Model" combines these predictions. Furthermore, "SharedCNNLayers" extract features for both modalities effectively, improving performance. This framework guarantees a thorough and forward-thinking method for multi-output emotion detection, in conjunction with code evolution and integration of Python libraries.

Dataset Description:

- This research integrated two independent datasets: the Mozilla Common Voice dataset (consists of 19,160 validated hours in 114 language-groups) for age-group and gender classification and the CREMA-D (which comprises the following emotions: Anger, Disgust, Fear, Happy, Neutral, and Sad) dataset for emotion recognition. The Mozilla Common Voice dataset has a broad range of voices from which we can extract features for age-group and gender prediction. It includes a large number of speakers, ensuring a diverse representation across demographic groups. The CREMA-D Dataset includes 7,442 audio snippets from 91 actors that were originally recorded. 2443 participants in the data collection process assessed the emotional content of the clips in three different modalities: audiovisual, video alone, and audio alone.

Table 1.



Category	Labels/Groups Targeted
Emotions	Anger, Disgust, Fear, Happy, Neutral, Sad
Genders	Male, Female (Binary)
Age Groups	Child, Young Adult, Adult, Middle Aged, Senior

- The predictive models' output labels are divided into three main groups: emotion, gender, and age-group. This three-pronged method allows for a more detailed understanding of voice-based elements and adds to a more comprehensive voice recognition system.
- Even though these separate datasets are extensive, at first there was trouble locating combined datasets that included labels for emotions, gender, and age-group Insufficient density or nonexistence of the given data made it impossible to train a reliable and accurate model. Direct integration was hindered

by the different data distribution between the public audio recordings in Common Voice and the actor-recorded emotional expressions in CREMA-D.

- To deal with this discrepancy, predictive models were trained independently on each dataset in order to capture the distinct features of CREMA-D and Common Voice. These models were then used to produce forecasts for a combined dataset. We used this synthesized dataset to train our Convolutional Neural Network (CNN) model, which combined predictions from both sources. The goal was to create a single model that could predict age-group, gender, and emotional states with accuracy.

Feature Extraction:

- This research feature extraction algorithm focuses on extracting various acoustic properties from audio recordings, providing useful insights for emotion, gender, and age-group recognition. Statistical measurements such as spectral centroid, bandwidth, rolloff, flatness, and contrast are among the retrieved features. These characteristics provide a thorough description of the audio signal, capturing both its central trends and spectrum qualities. The interpretability, computational efficiency, and relevance to voice-based recognition tasks drive the selection of characteristics.
- Because different groups exhibit varying frequency characteristics, the spectral centroid, a measure of average-group frequency, is susceptible to fluctuations in emotion, gender, and age-group. Spectral bandwidth, which indicates the frequency spread, and spectral rolloff, which defines the frequency below which a certain proportion of energy lies, help to capture changes in high-frequency content, which can be indicative of different emotional states, genders, and age-group groups.
- Mel-Frequency Cepstral Coefficients (MFCCs) have been added to the feature extraction procedure to enhance it. MelFrequency Cepstral Coefficients provide a useful frequency domain representation of the audio signal's short-term power spectrum. We have customized the dataset by excluding MFCC12 and MFCC-19.

Implementation and Source code:

A. Approach 1:Individual Modal

- Emotion:
  An LSTM neural network was trained to recognise emotions using data from RAVDESS, CREMAD, Tess, and Savee datasets. MFCCs, Zero Crossing Rate,

and Root Mean Square Energy were among the features that captured audio spectral and temporal characteristics. Sequential LSTM layers captured temporal dependencies, with feature input layers. Overfitting was avoided by using dropout layers. The softmax-activated output layer predicted emotion classes.Categorical cross entropy loss gauged performance, appropriate for multi-class jobs. RMSProp was optimized, and categorical accuracy was utilized for training. Learning rates were changed via a ReduceLROnPlateau callback. The model was trained for 100 epochs, with batch size. Validation accuracy was 90%.

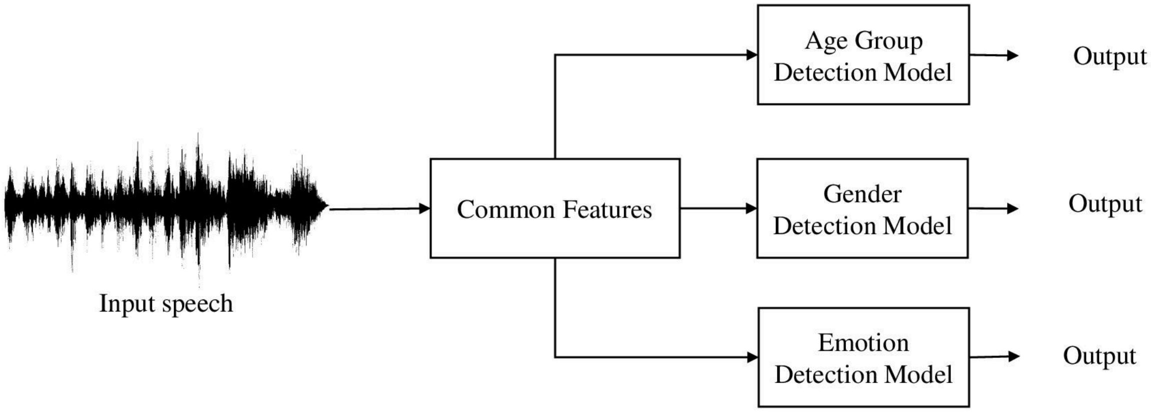


Fig 2.

- Gender:
  The gender prediction model uses a compact CNN architecture tailored for binary gender classification from speech signals. This network has 2 convolutional layers, max pooling layers and 2 fully connected layers with dropout regularization. The output layer predicts male or female gender using a sigmoid activation. Binary cross entropy loss and the RMSprop optimizer were used. After training for 30 epochs and using early stopping, the model obtained 89.4% accuracy in categorizing speaker gender on the unseen test set. Precision was measured at 0.91 for the male category and 0.86 for the female category. The confusion matrix showed a small skew towards more male gender predictions overall. Investigation showed pitch-based features as providing greater distinguishing evidence for gender than spectral features such as MFCCs.

- Age Group:

The age-group prediction model utilizes a convolutional neural network (CNN) architecture optimized for multi-class classification, predicting the speaker's age-group category from speech. The model contains 3 convolutional layers interweaved with 2 max pooling layers, followed by 2 fully connected layers of 128 and 64 units respectively. ReLU activation and batch normalization were utilized between layers. The model was trained using the categorical cross entropy loss function along with the Adam optimizer, with a learning rate of 0.001 for 100 epochs and a batch size of 32. The Common Voice dataset was split into 80% training, 10% validation, and 10% test subsets. Training samples were randomly augmented via time shifting and background noise injection. The model achieved an overall accuracy of 71.25% on the age-group category classification task, with a precision of 0.74 and recall of 0.69 on the test data. Particularly strong performance was noted in young adult and middle-age-group groups, while weaker accuracy was achieved in senior age-group bands above 60 years. Misclassifications tended to occur most often between adjacent age-group categories.

B. Approach 2 : Age-Group,Gender in Single model followed by Emotion model

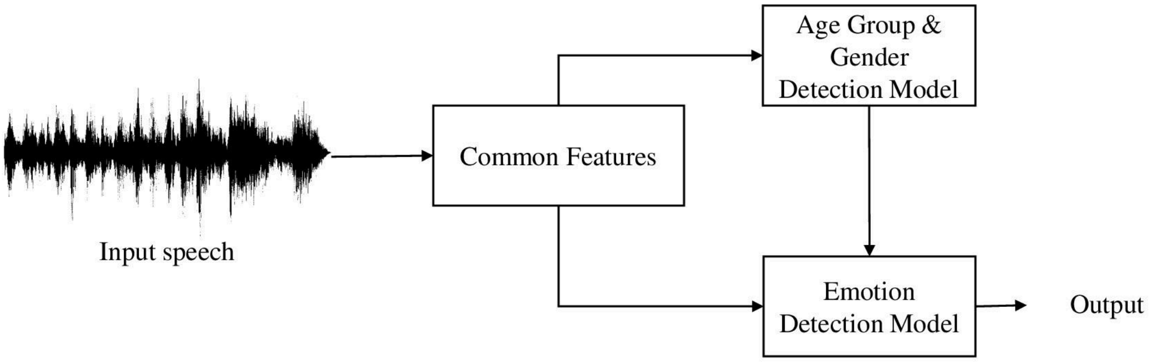


Fig 3.

- The research uses the Mozilla Common Voice dataset and a unified model to investigate age-group and gender recognition. During preprocessing, pertinent features and labels pertaining to audio frequency characteristics, gender, and age-group are extracted. A thorough representation of the audio data is provided by the extracted features, which include the spectral centroid, bandwidth, rolloff, and Mel-Frequency Cepstral Coefficients (MFCCs).

- Feature scaling is used to increase the robustness of the model, and the ANOVA statistical method is employed for feature selection. K-Fold Cross-Validation is used to train and assess two classifiers, Support Vector Machine (SVM) and Random Forest, with the F1-Score serving as a crucial performance indicator. Through an internal Cross-Validation method, classifier hyperparameters are optimized.
- The SVM classifier outperformed the Random Forest classifier with an accuracy of 82.7%, which is promising compared to its 71.8% .

C. Approach 3: Gender,Emotion in Single model followed by age group group model

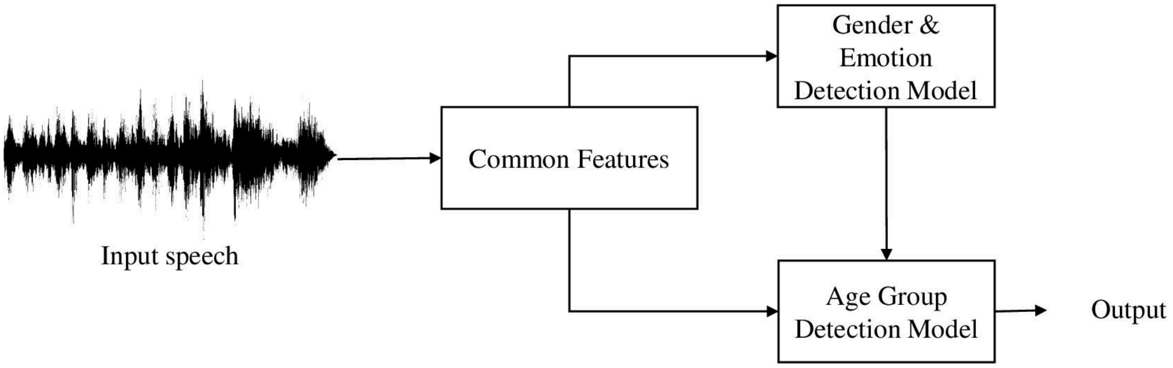


Fig 4.

- The research used the RAVDESS dataset, which contains labeled audio for both gender and emotion recognition. Emotions were categorized as neutral, happy, sad, and furious. The dataset underwent partitioning into training and testing sets. Key speech signal features, including Mel-Frequency Cepstral Coefficients (MFCCs), Zero Crossing Rate, and Root Mean Square Energy, were extracted using Wav2Vec2FeatureExtractor from transformers.
- HubertForSequenceClassification, the model architecture adopted, is a pre-trained model fine-tuned for sequence classification. Training consisted of two epochs, each of which processed batches of size 2 through a DataLoader. The optimizer was Adam, with a learning rate of 1e-5. The evaluation metric chosen was categorical accuracy. On the test dataset, the training phase produced a considerable accuracy of 74.57%, indicating the model's ability to distinguish

gender and emotion from speech data. This demonstrates its potential for real-world applications needing in-depth voice analysis.

D. Approach 4: Unified model age-group group, Gender and Emotion

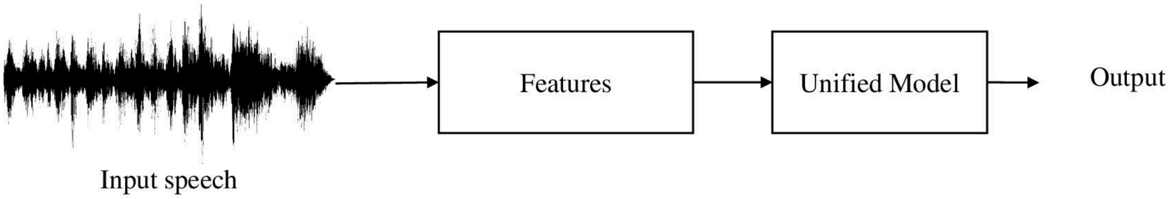


Fig 5.

- Using CREMA-D for emotion and Mozilla Common Voice for gender and age-group, the research explores a combination model for age-group, gender, and emotion recognition. The dataset is processed using encoding and one-hot encoding for the labels of emotion, age-group, and gender in order to separate features and labels. The dataset is divided, and then the input features are standardized.
- This approach aligns seamlessly with our overarching research goal of establishing a unified paradigm for comprehensive voice-based recognition. The findings highlight that simultaneous age-group and gender classification significantly improves performance in contrast to single-task models. Various speaker datasets are employed for age-group and gender classification in the research, resonating with the broader scope of our exploration into unified model research.
- Convolutional Neural Network (CNN) layers for feature extraction and later dense layers for abstraction are integrated into the model architecture. The output layers— emotion, age-group, and gender—are customized for every recognition task. For each task, a categorical cross entropy loss version of the model is compiled and then optimized using the Adam optimizer. Fitting the model to training data and evaluating it on a test set comprises training.
- On the test dataset, the model shows 53.72% accuracy in predicting emotion, 41.77% accuracy in predicting age group, and 48.09% accuracy in predicting gender.

A variety of metrics were used to evaluate the model's performance, and one such metric is the F1 score, which offers a balanced measurement based on recall and precision.

Recall computes the ratio of genuine positives to real positives, whereas precision measures the accuracy of positive forecasts. The F1 score provides a single accuracy statistic by balancing precision and recall by incorporating erroneous positives and false negatives. Standard criteria such as accuracy, precision, and recall were used in addition to the F1 score to assess the model's efficacy. Accuracy offers a comprehensive assessment of overall performance, memory records pertinent positive examples, and precision measures accurate positive predictions.

Below is the source code for the unified model:

~~~
```


import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder
from keras.models import Model
from keras.layers import Input, Conv1D, MaxPooling1D, Flatten, Dense,
Concatenate
from keras.optimizers import Adam
from keras.utils import to_categorical

# Load the CSV file
df = pd.read_csv('edited_file.csv')

# Assuming the features are in columns 0 to 20, and the labels are in
columns 21, 22, and 23
X = df.iloc[:, :21].values
y_emotion = df.iloc[:, 21].values
y_age = df.iloc[:, 22].values
y_gender = df.iloc[:, 23].values

# Use LabelEncoder for 'emotion'
label_encoder = LabelEncoder()
y_emotion_encoded = label_encoder.fit_transform(y_emotion)
num_emotion_classes = len(label_encoder.classes_)

# Convert categorical labels to one-hot encoding
y_emotion = to_categorical(y_emotion_encoded,
num_classes=num_emotion_classes)
y_age = to_categorical(y_age, num_classes=8)
y_gender = to_categorical(y_gender, num_classes=2)

# Split the data into training and testing sets
~~~

```
X_train, X_test, y_emotion_train, y_emotion_test, y_age_train, y_age_test,
y_gender_train, y_gender_test = train_test_split(
    X, y_emotion, y_age, y_gender, test_size=0.2, random_state=42
)

# Standardize the input features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# CNN Layers
inputs_cnn = Input(shape=(21, 1))
conv1 = Conv1D(32, 3, activation='relu')(inputs_cnn)
pool1 = MaxPooling1D(pool_size=2)(conv1)
conv2 = Conv1D(64, 3, activation='relu')(pool1)
pool2 = MaxPooling1D(pool_size=2)(conv2)
flat_cnn = Flatten()(pool2)

# Dense Layers
dense1 = Dense(64, activation='relu')(flat_cnn)
dense2 = Dense(64, activation='relu')(flat_cnn)
dense3 = Dense(64, activation='relu')(flat_cnn)

# Output Layers
output_emotion = Dense(num_emotion_classes, activation='softmax',
name='emotion')(dense1)
output_age = Dense(8, activation='softmax', name='age')(dense2)
output_gender = Dense(2, activation='softmax', name='gender')(dense3)

# Create the model
model = Model(inputs=inputs_cnn, outputs=[output_emotion, output_age,
output_gender])

# Compile the model
model.compile(optimizer=Adam(lr=0.001),
              loss={'emotion': 'categorical_crossentropy', 'age':
'categorical_crossentropy', 'gender': 'categorical_crossentropy'},
              metrics={'emotion': 'accuracy', 'age': 'accuracy', 'gender':
'accuracy'})

# Display the model summary
model.summary()

# Train the model
model.fit(X_train, [y_emotion_train, y_age_train, y_gender_train],
epochs=10, batch_size=32, validation_split=0.2)
```

```
__________________________________________________________________________________________
 Layer (type)                Output Shape                 Param #   Connected to
==================================================================================================
 input_1 (InputLayer)        [(None, 21, 1)]              0         []

 conv1d (Conv1D)             (None, 19, 32)               128       ['input_1[0][0]']

 max_pooling1d (MaxPooling1  (None, 9, 32)                0         ['conv1d[0][0]']
 D)

 conv1d_1 (Conv1D)           (None, 7, 64)                6208      ['max_pooling1d[0][0]']

 max_pooling1d_1 (MaxPoolin  (None, 3, 64)                0         ['conv1d_1[0][0]']
 g1D)

 flatten (Flatten)           (None, 192)                  0         ['max_pooling1d_1[0][0]']

 dense (Dense)               (None, 64)                   12352     ['flatten[0][0]']

 dense_1 (Dense)             (None, 64)                   12352     ['flatten[0][0]']

 dense_2 (Dense)             (None, 64)                   12352     ['flatten[0][0]']

 emotion (Dense)             (None, 6)                    390       ['dense[0][0]']

 age (Dense)                 (None, 8)                    520       ['dense_1[0][0]']

 gender (Dense)              (None, 2)                    130       ['dense_2[0][0]']

==================================================================================================
Total params: 44432 (173.56 KB)
Trainable params: 44432 (173.56 KB)
Non-trainable params: 0 (0.00 Byte)
__________________________________________________________________________________________
Epoch 1/10
149/149 [==============================] - 5s 13ms/step - loss: 3.8212 - emotion_loss:
1.5492 - age_loss: 1.6518 - gender_loss: 0.6201 - emotion_accuracy: 0.3578 - age_accuracy:
0.3566 - gender_accuracy: 0.6697 - val_loss: 3.6718 - val_emotion_loss: 1.5046 - val_age_loss:
1.5656 - val_gender_loss: 0.6017 - val_emotion_accuracy: 0.3829 - val_age_accuracy: 0.3728 -
val_gender_accuracy: 0.6692
Epoch 2/10
```

```
149/149 [==============================] - 1s 6ms/step - loss: 3.5782 - emotion_loss:
1.4523 - age_loss: 1.5407 - gender_loss: 0.5852 - emotion_accuracy: 0.4072 - age_accuracy:
0.3727 - gender_accuracy: 0.7016 - val_loss: 3.5674 - val_emotion_loss: 1.4719 - val_age_loss:
1.5062 - val_gender_loss: 0.5894 - val_emotion_accuracy: 0.4022 - val_age_accuracy: 0.4240 -
val_gender_accuracy: 0.6902
Epoch 3/10
149/149 [==============================] - 1s 6ms/step - loss: 3.4522 - emotion_loss:
1.4120 - age_loss: 1.4748 - gender_loss: 0.5654 - emotion_accuracy: 0.4286 - age_accuracy:
0.4166 - gender_accuracy: 0.7199 - val_loss: 3.5079 - val_emotion_loss: 1.4548 - val_age_loss:
1.4632 - val_gender_loss: 0.5899 - val_emotion_accuracy: 0.4039 - val_age_accuracy: 0.4500 -
val_gender_accuracy: 0.6851
Epoch 4/10
149/149 [==============================] - 1s 5ms/step - loss: 3.3703 - emotion_loss:
1.3932 - age_loss: 1.4277 - gender_loss: 0.5495 - emotion_accuracy: 0.4360 - age_accuracy:
0.4414 - gender_accuracy: 0.7274 - val_loss: 3.4796 - val_emotion_loss: 1.4482 - val_age_loss:
1.4522 - val_gender_loss: 0.5792 - val_emotion_accuracy: 0.4123 - val_age_accuracy: 0.4408 -
val_gender_accuracy: 0.6885
Epoch 5/10
149/149 [==============================] - 1s 5ms/step - loss: 3.3046 - emotion_loss:
1.3726 - age_loss: 1.3863 - gender_loss: 0.5457 - emotion_accuracy: 0.4462 - age_accuracy:
0.4551 - gender_accuracy: 0.7306 - val_loss: 3.4278 - val_emotion_loss: 1.4245 - val_age_loss:
1.4135 - val_gender_loss: 0.5898 - val_emotion_accuracy: 0.4139 - val_age_accuracy: 0.4568 -
val_gender_accuracy: 0.7011
Epoch 6/10
149/149 [==============================] - 1s 5ms/step - loss: 3.2385 - emotion_loss:
1.3496 - age_loss: 1.3535 - gender_loss: 0.5354 - emotion_accuracy: 0.4584 - age_accuracy:
0.4675 - gender_accuracy: 0.7379 - val_loss: 3.3745 - val_emotion_loss: 1.4144 - val_age_loss:
1.3844 - val_gender_loss: 0.5757 - val_emotion_accuracy: 0.4274 - val_age_accuracy: 0.4694 -
val_gender_accuracy: 0.7011
Epoch 7/10
149/149 [==============================] - 1s 6ms/step - loss: 3.1808 - emotion_loss:
1.3326 - age_loss: 1.3202 - gender_loss: 0.5281 - emotion_accuracy: 0.4624 - age_accuracy:
0.4859 - gender_accuracy: 0.7415 - val_loss: 3.3816 - val_emotion_loss: 1.4251 - val_age_loss:
1.3927 - val_gender_loss: 0.5638 - val_emotion_accuracy: 0.4316 - val_age_accuracy: 0.4593 -
val_gender_accuracy: 0.7112
Epoch 8/10
149/149 [==============================] - 1s 6ms/step - loss: 3.1294 - emotion_loss:
1.3225 - age_loss: 1.2902 - gender_loss: 0.5168 - emotion_accuracy: 0.4660 - age_accuracy:
0.5010 - gender_accuracy: 0.7522 - val_loss: 3.3483 - val_emotion_loss: 1.4081 - val_age_loss:
1.3563 - val_gender_loss: 0.5839 - val_emotion_accuracy: 0.4307 - val_age_accuracy: 0.4727 -
val_gender_accuracy: 0.7053
Epoch 9/10
149/149 [==============================] - 1s 6ms/step - loss: 3.0790 - emotion_loss:
1.3033 - age_loss: 1.2620 - gender_loss: 0.5137 - emotion_accuracy: 0.4796 - age_accuracy:
0.4990 - gender_accuracy: 0.7488 - val_loss: 3.3404 - val_emotion_loss: 1.4082 - val_age_loss:
```

```
1.3448 - val_gender_loss: 0.5874 - val_emotion_accuracy: 0.4215 - val_age_accuracy: 0.4744 -
val_gender_accuracy: 0.6910
Epoch 10/10
149/149 [==============================] - 1s 5ms/step - loss: 3.0254 - emotion_loss:
1.2908 - age_loss: 1.2293 - gender_loss: 0.5052 - emotion_accuracy: 0.4813 - age_accuracy:
0.5185 - gender_accuracy: 0.7583 - val_loss: 3.3541 - val_emotion_loss: 1.4164 - val_age_loss:
1.3560 - val_gender_loss: 0.5817 - val_emotion_accuracy: 0.4316 - val_age_accuracy: 0.4660 -
val_gender_accuracy: 0.6961
```

Out[3]:

```
<keras.src.callbacks.History at 0x7fac846f7a60>
```

In [4]:

```
# Evaluate the model on the test set
evaluation = model.evaluate(X_test, [y_emotion_test, y_age_test,
y_gender_test])

# Print the evaluation metrics
print(f"Test Loss: {evaluation[0]}")
print(f"Emotion Accuracy: {evaluation[3] * 100:.2f}%")
print(f"Age Accuracy: {evaluation[4] * 100:.2f}%")
print(f"Gender Accuracy: {evaluation[5] * 100:.2f}%")
```

```
47/47 [==============================] - 0s 3ms/step - loss: 3.3044 - emotion_loss: 1.4292 -
age_loss: 1.3380 - gender_loss: 0.5372 - emotion_accuracy: 0.4177 - age_accuracy: 0.4809 -
gender_accuracy: 0.7314
Test Loss: 3.304385185241699
Emotion Accuracy: 53.72%
Age Accuracy: 41.77%
Gender Accuracy: 48.09%
```

In [8]:

```
import numpy as np

# Custom input
custom_input = np.array([[2679.9395691628692 ,3347.669488765762,
5745.486745886655
 -625.2181396484375, 111.32093811035156 ,6.3269944190979
,34.75761413574219,
 31.619901657104492 ,-4.714645862579346 ,-0.486030638217926,
 -4.934024333953857 ,-12.714733123779297 ,-2.0551483631134033,
 -3.7411177158355713, -10.702962875366211, -11.20263671875,
 -12.003522872924805, -8.489580154418945 ,-5.463275909423828,
```

~~~
 -4.954216480255127, -3.715198278427124,-6.338474273681641]])

# Standardize the custom input
custom_input = scaler.transform(custom_input)

# Reshape the input to match the model's input shape
custom_input = custom_input.reshape((1, 21, 1))

# Make predictions
predictions = model.predict(custom_input)

# Extract the predictions for each output
emotion_prediction = predictions[0]
age_prediction = predictions[1]
gender_prediction = predictions[2]

# Decode emotion predictions using the LabelEncoder
emotion_prediction_decoded =
label_encoder.inverse_transform(np.argmax(emotion_prediction, axis=1))

# Print the predictions
print(f"Emotion Prediction: {emotion_prediction_decoded}")
print(f"Age Prediction: {np.argmax(age_prediction, axis=1)}")
print(f"Gender Prediction: {np.argmax(gender_prediction, axis=1)}")
~~~

```
1/1 [==============================] - 0s 35ms/step
Emotion Prediction: [0]
Age Prediction: [7]
Gender Prediction: [0]
```

In [ ]:

~~~
# Save the entire model to a HDF5 file
model.save('all_model.h5')

```
~~~

The emotions predicted are as follows:
Fear : 0, Anger: 1, Disappointment: 2, Sad: 3, Neutral: 4, Happy: 5.

The Age groups are as follows:

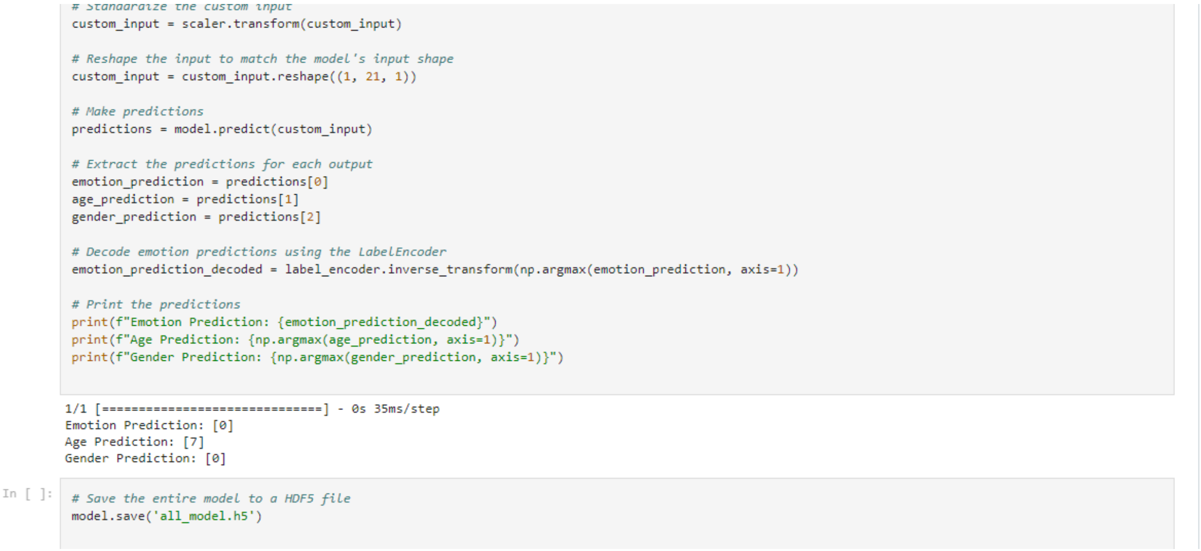
1. Teens: '< 19'
2. Twenties: '19 - 29'

3. Thirties: '30 - 39'
4. Fourties: '40 - 49'
5. Fifties: '50 - 59'
6. Sixties: '60 - 69'
7. Seventies: '70 - 79'
8. Eighties: '80 - 89'
9. Nineties: '> 89'

The genders are as follows:

1. Male, 2. Female, 3. Other

Screenshots of the solution:

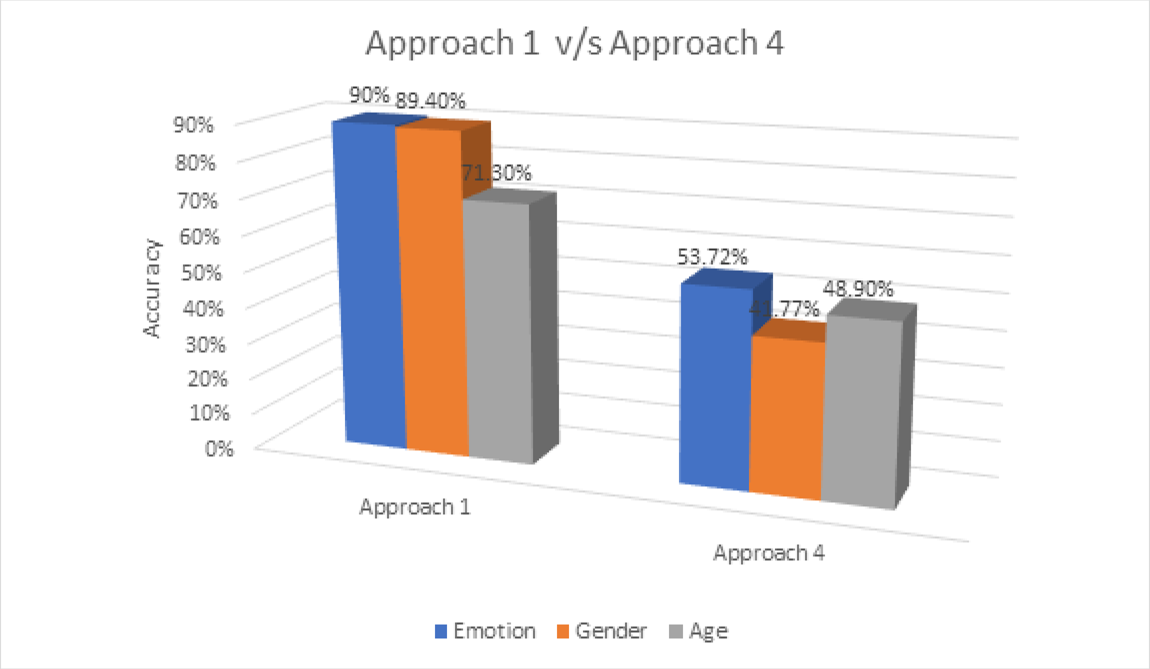


```
# Standardize the custom input
custom_input = scaler.transform(custom_input)

# Reshape the input to match the model's input shape
custom_input = custom_input.reshape((1, 21, 1))

# Make predictions
predictions = model.predict(custom_input)

# Extract the predictions for each output
emotion_prediction = predictions[0]
age_prediction = predictions[1]
gender_prediction = predictions[2]

# Decode emotion predictions using the LabelEncoder
emotion_prediction_decoded = label_encoder.inverse_transform(np.argmax(emotion_prediction, axis=1))

# Print the predictions
print(f"Emotion Prediction: {emotion_prediction_decoded}")
print(f"Age Prediction: {np.argmax(age_prediction, axis=1)}")
print(f"Gender Prediction: {np.argmax(gender_prediction, axis=1)}")
```

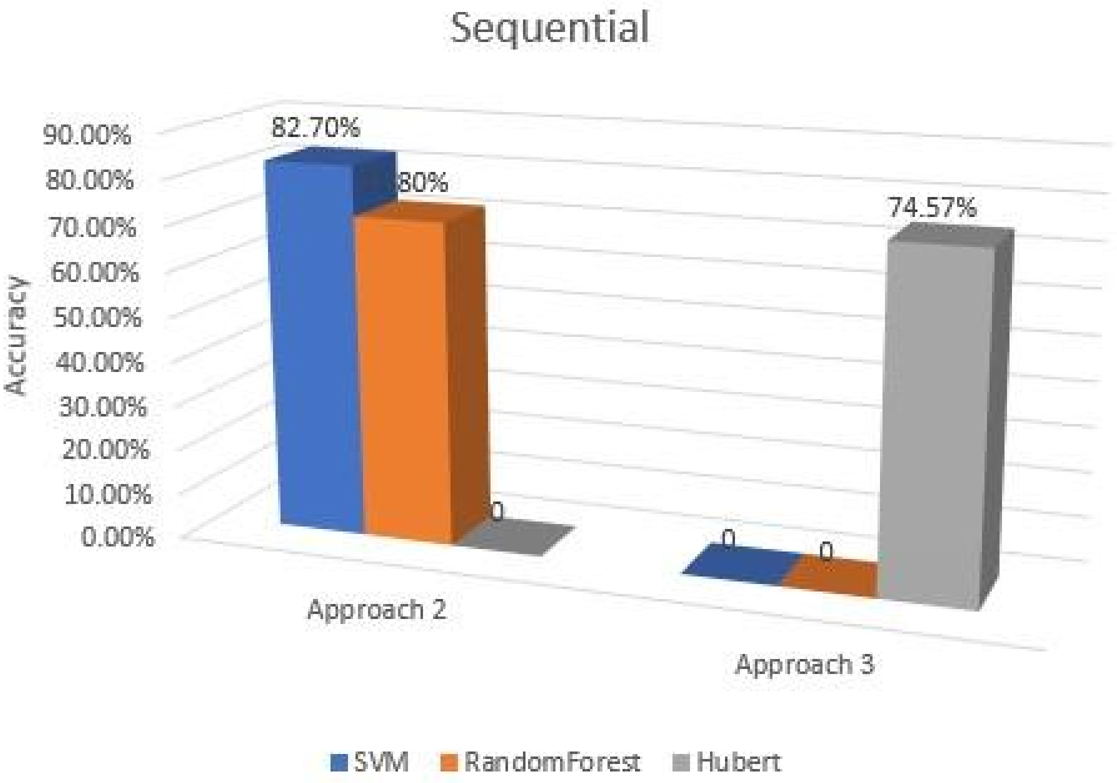
```
1/1 [==============================] - 0s 35ms/step
Emotion Prediction: [0]
Age Prediction: [7]
Gender Prediction: [0]
```

```
In [ ]: # Save the entire model to a HDF5 file
        model.save('all_model.h5')
```

Fig 6.


Approach 1 v/s Approach 4
Accuracy
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
90%
89.40%
71.30%
53.72%
41.77%
48.90%
Approach 1
Approach 4
Emotion
Gender
Age

Graph 1.


Sequential
Accuracy
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
82.70%
80%
0
74.57%
0
0
Approach 2
Approach 3
SVM
RandomForest
Hubert

Graph 2.

Conclusion:

In summary, our research has made significant strides in the realm of voice analysis, particularly in the domains of gender, age-group, and emotion recognition. The research's outcomes underscored the effectiveness of our integrated models in discerning these diverse features within audio data. Notably, the results highlighted the pivotal role of tailored model selection, with the Support Vector Machine (SVM) outperforming the Random Forest classifier in age group and gender detection. The gender and emotion identification model showcased the capability to extract intricate aspects from speech data with commendable accuracy levels. Encouragingly, the unified model addressing age-group, gender, and emotion detection demonstrated promising results, suggesting potential applications in scenarios requiring comprehensive voice analysis. Looking ahead, there exists ample room for future endeavors. Possible enhancements to current models could involve fine-tuning hyperparameters, exploring more intricate neural network structures, and incorporating additional datasets to enhance diversity. Investigating transfer learning strategies, where models pretrained on extensive datasets are customized for our specific needs, holds promise for performance improvement. Additionally, addressing potential biases in datasets and refining preprocessing methods can contribute to the creation of more reliable and unbiased models. Future research directions may include delving into real-time applications and adapting models for implementation in diverse contexts. Ultimately, our research lays a robust foundation, and subsequent projects can leverage-group these findings to propel the capabilities of voice analysis systems for a myriad of practical applications.

References:

[1]A. Mehrish, N. Majumder, R. Bhardwaj, R. Mihalcea, and S. Poria, "A Review of Deep Learning Techniques for Speech Processing," (2023).

[2] S. R. Zaman, D. Sadekeen, M. A. Alfaz and R. Shahriyar, "One Source to Detect them All: Gender, age-group, and Emotion Detection from Voice," 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 2021, pp. 338-343, doi: 10.1109/COMPSAC51774.2021.00055.

[3] Sánchez-Hevia, H.A., Gil-Pita, R., Utrilla-Manso, M. et al. age-group group classification and gender recognition from speech with temporal convolutional neural networks. Multimed Tools Appl 81, 3535–3552 (2022)

[4] Poonam Rani et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 4, Dec 2018, pp. 14-17.

[5] Jinkyu Lee, Ivan Tashev, High Level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition (2015).

[6] Gómez García, Jorge Andrés , Moro Velázquez, Laureano, Godino Llorente, Juan Ignacio and Castellanos Domínguez, Germán (2015). Automatic age-group detection in normal and pathological voice. In: "16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)", 6/09/2015 - 10/09/2015, Dresden (Germany). ISBN a. 978-1-5108-1790-6. pp. 3739-3743.

[7] Prasanta Roy, Parabattina Bhagath, and Pradip Das. 2020. Gender Detection from Human Voice Using Tensor Analysis. In Proceedings of the 1st Joint Workshop on Spoken Language-group Technologies for Underresourced language-groups (SLTU).

[8] Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z., Li, R. (2020) Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. Proc. Interspeech 2020, 394-398, doi: 10.21437/Interspeech.2020-1705 .

[9] Scikit-Learn Classification Metrics Documentation.

[10] Poonam Rani et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 4, Dec 2018, pp. 14- 17.

[11] Jinkyu Lee, Ivan Tashev, High Level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition (2015).

[12] Gómez García, Jorge Andrés , Moro Velázquez, Laureano, Godino Llorente, Juan Ignacio and Castellanos Domínguez, Germán (2015). Automatic age-group detection in normal and pathological voice. In: "16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)", 6/09/2015 - 10/09/2015, Dresden (Germany). ISBN [13] 978-1-5108-1790-6. pp. 3739-3743.

[14] Prasanta Roy, Parabattina Bhagath, and Pradip Das. 2020. Gender Detection from Human Voice Using Tensor Analysis. In Proceedings of the 1st Joint Workshop on Spoken Language-group Technologies for Under-resourced language-groups (SLTU).

[15] Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z., Li, R. (2020) Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. Proc. Interspeech 2020, 394-398, doi: 10.21437/Interspeech.2020-1705 .

[16] Scikit-Learn Classification Metrics Documentation.

[17] Spectral Centroid:The spectral centroid is a measure used in digital signal processing to characterize a spectrum.Bandwidth:Bandwidth specifically refers to the capacity at which a network can transmit data.Flatness:Flatness is a soft, short tone heard when percussing over solid tissue like muscle and bone. Contrast:small differences in speech sounds, that makes a difference in how the sound is perceived by listeners.

[18] In sound processing, the mel-frequency cepstral Coefficient (MFCC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. [19] age-group and emotion:: S. R. Zaman, D. Sadekeen, M. A. Alfaz and R. Shahriyar.

[20] emotion:Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, Rongjun Li.

[21] emotion:Jinkyu Lee, Ivan Tashev.