**Methodology**

This section describes our strategy of employing Convolutional Neural Networks (CNNs) to solve the problems of age prediction, gender classification, and emotion identification using audio data. We propose a thorough technique that includes developing different CNN models adapted to each distinct job, investigating sequential models to capture temporal dependencies, and developing a single multimodal model for a more holistic analysis.

Individual Models:

We suggested the building of three unique CNN models to establish a firm baseline and get insights into the performance of each modality:

Age Prediction Model: Based on the speaker's voice, this model predicts their age. With an architecture optimized for age prediction, it has numerous convolutional layers, pooling layers, and fully linked layers.

Gender Prediction Model: Based on the speaker's speech, this model predicts whether they are male or female. It has convolutional layers, pooling layers, and fully linked layers optimized to gender categorization, just like the age prediction model.

Emotion Prediction Model: This model determines the speaker's emotion among categories such as anger, disgust, fear, happiness, sadness or neutral. The architecture has been optimized for emotion recognition, with convolutional, pooling, and fully linked layers included.

Sequential Models:

We suggested the investigation of two alternative designs in order to capture the temporal relationships that may exist between age, gender, and emotion:

Early Fusion: In this design, the outputs of the different CNN models were concatenated and fed into a final fully connected layer for joint prediction. We captured the interdependencies between tasks by combining information from these models early in the process.

Late Fusion: Individual CNN models developed their own forecasts under this architecture. To aggregate these predictions and determine the final conclusion, a fusion layer was used. Late fusion enabled a more adaptable and modular methodology, allowing predictions to be included at a later stage.

Single Multimodal Model: Our goal was to conduct a more comprehensive analysis that captured the interdependence of age, gender, and emotion. We tried creating a single multimodal model. This algorithm immediately analyzed the audio signal while also predicting the speaker's age, gender, and sentiment. It used a shared convolutional layer to extract features from audio signals, followed by independent branches for each prediction task.

**Result**

In order to detect age, gender, and emotion from voice, our project used both separate models and a single multimodal model, each of which was adapted to certain tasks. The results of these models provide information about their capabilities and overall performance.

Age Prediction Model: For age prediction, we used a thick Artificial Neural Network (ANN). This model had an accuracy of 94.15%. The model was trained on the Mozilla Common Voice dataset and predicted age categories such as teenagers, twenties, thirties, forties, fifties, sixties, seventies, and eighties with success. It proved its worth with a batch size of 256 and training over 26 epochs.

Gender Prediction Model: The gender prediction model, implemented as a Neural Network using MLPClassifier, achieved a high accuracy of 98%. The model was trained on Kaggle's "Gender Recognition by Voice" dataset, demonstrating its ability to reliably classify speakers based on their gender.

Emotion Prediction Model: We used Neural Networks, MLPClassifier to get a noteworthy accuracy of 77.6%. It effectively identified a spectrum of emotions in the speakers' voices after being trained on the RAVDESS dataset, highlighting its competency in this particular task.

Multimodal Model in One: A holistic approach to voice analysis was offered by a single multimodal model that combined age, gender, and emotion prediction inside a unified architecture. It processed audio signals effectively using a CNN model. This model performed well in our testing, with an overall accuracy of 53.41% for emotion detection, 41.71% for age prediction, and 46.41% for gender categorization.

Crema D and Mozilla Common Voice datasets were used in the study, which contributed to the varied range of sounds and features. The combination of these distinct models with the unified multimodal model provides a complete solution for voice-based age, gender, and emotion recognition, allowing for a more sophisticated understanding of the speaker's characteristics and emotional states.

**Conclusion**

The results of this project demonstrate both the capabilities of our models and the possibilities for future advances in this sector.

Our proposed model's merits are evident in its multidimensional approach. Individual models suited to each job, such as age prediction, gender classification, and emotion identification, perform exceptionally well, with age and gender prediction reaching up to 98% and 94.15%, respectively. This highlights the value of using specialized models for each attribute. Furthermore, our single multimodal model, which is designed to process age, gender, and emotion simultaneously, provides a holistic analysis of voice data, providing a balanced accuracy of 53.41% for emotion, 41.71% for age, and 46.41% for gender. This method not only demonstrates the possibility of a unified understanding of speaker characteristics, but it also sets the groundwork for a more comprehensive voice analysis framework.

While our initiative has shown encouraging results, there are various opportunities for further research and development. First, the creation of more advanced multimodal models that combine diverse data sources could improve the system's accuracy and robustness. Second, broadening the datasets used to include a greater range of age groups, genders, and emotional expressions can result in more comprehensive and generalizable models. Incorporating advanced feature engineering approaches and investigating cutting-edge deep learning architectures may also lead to improved predictive performance. Furthermore, real-world implementations of these models, such as in the fields of human-computer interaction and emotional computing, require extensive research to determine their practical utility.