Age Prediction through Face & Voice

Team 7 - Shwetha Iyer + Radhi Priya Janakiraman + Nada Khan + Chavi Mangla + Rukmini Manojkumar

Outline

Motivation

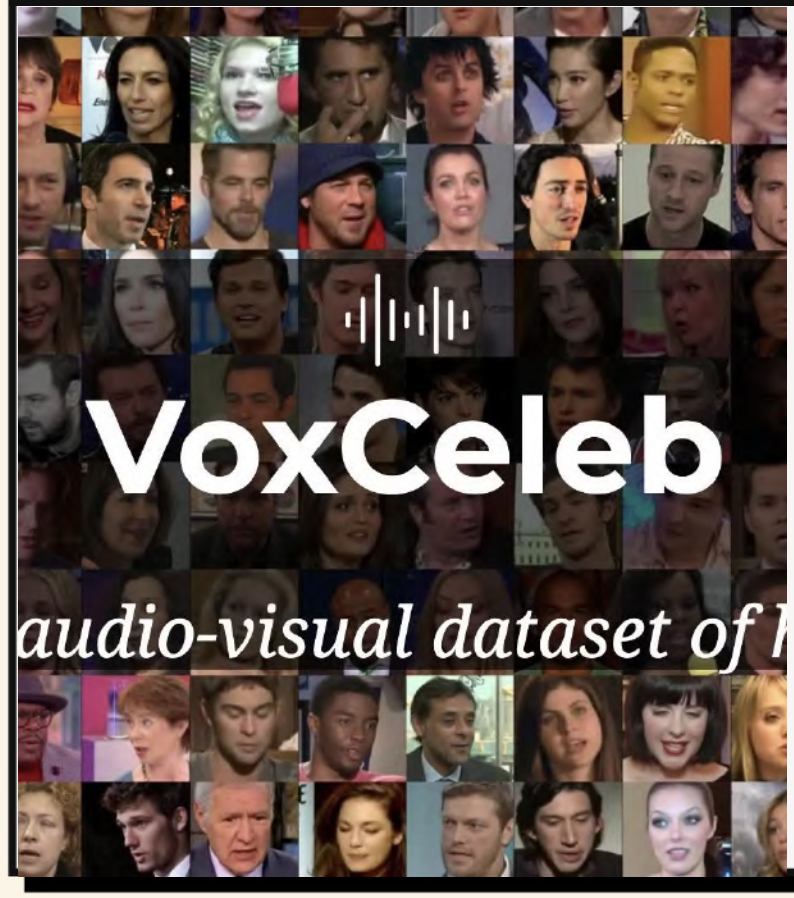
In the realm of online safety, the pressing issue revolves around implementing **reliable age verification** mechanisms through video verification to **prevent underage users from accessing potentially harmful or inappropriate content**, thus fostering a safer digital environment for individuals of varying age groups.

Datasets

Data

1. Problem
2. Image Datasets
3. Audio Datasets
3. Sampling Distribution

Problem with Video Datasets



Age VoxCeleb

The AgeVoxCeleb contains nearly 168k videos of approximately 5000 speakers. All the videos are labeled with the speaker’s real age estimated using each celebrity's name and title of the original YouTube video.

Problems

- Missing Data (content hosted on Youtube)
- Noisy audio
- Low quality images
- Multiple speakers
- Unsure of which speaker the age label belongs to

Audio Datasets

CREMA-D

An audio-visual dataset for emotion recognition containing 7,442 clips of 91 actors from diverse ethnic backgrounds.

The Eugene Children's Story Corpus (ECSC)

Includes 367 audio recordings and transcriptions of structured spontaneous narratives elicited from a total of 188 typically developing school-aged children.

English Children

The dataset contains audio recordings (lossless WAV) of 11 young children (age M=4.9 years old; 5 females, 6 males).

audioMNIST

The dataset consists of 30,000 audio samples of spoken digits (0 - 9). A subset of 9,623 samples was used.

Speech Accent

An audio dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube.

Image Datasets

UTKFace

Consists of over 20,000 face images with annotations of age (long age span), gender and ethnicity. The images cover large variations in poses, facial expression, resolution etc.

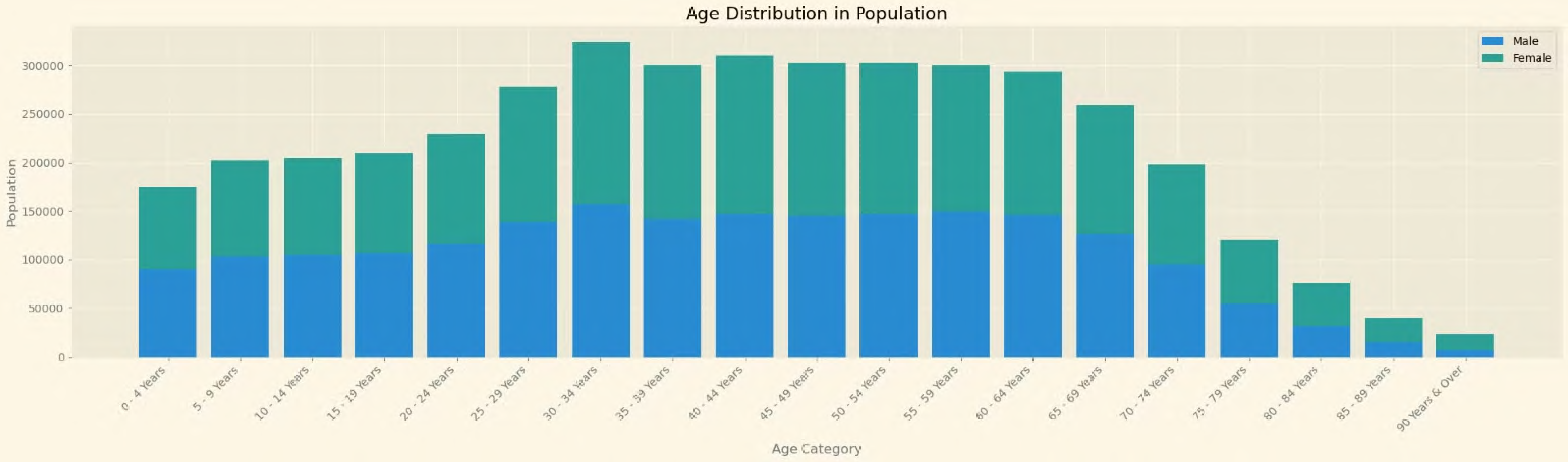
All-Age-Faces

Contains 13,322 face images (mostly Asian) distributed across all ages (from 2-80) including 7381 females and 5941 males.

Initial Dataset Distribution

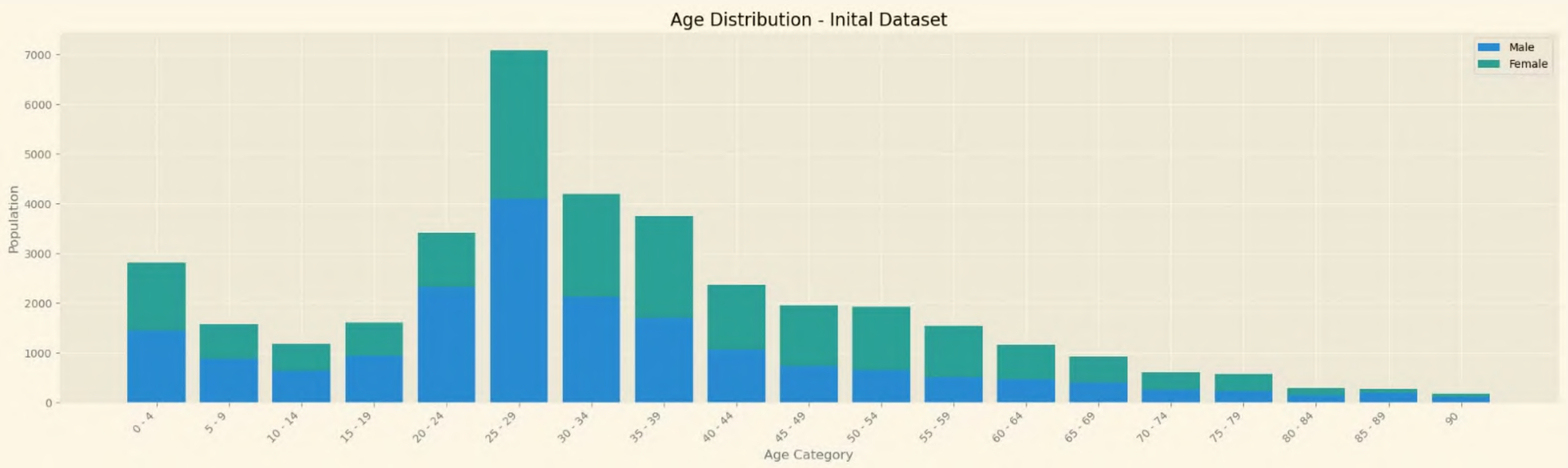
Source	Age	Number of Samples	Input Type
CREMA-D	20 - 79	7,442	Audio
AudioMNIST	20 - 69	9,623	Audio
Speech Accent Kaggle	12 - 80	2,138	Audio
english_children	<12	342	Audio
CHILDES Frogs ECSC	<12	341	Audio
UTKFace	0 - 116	21,205	Image
All-Age-Faces	2-82	13,322	Image

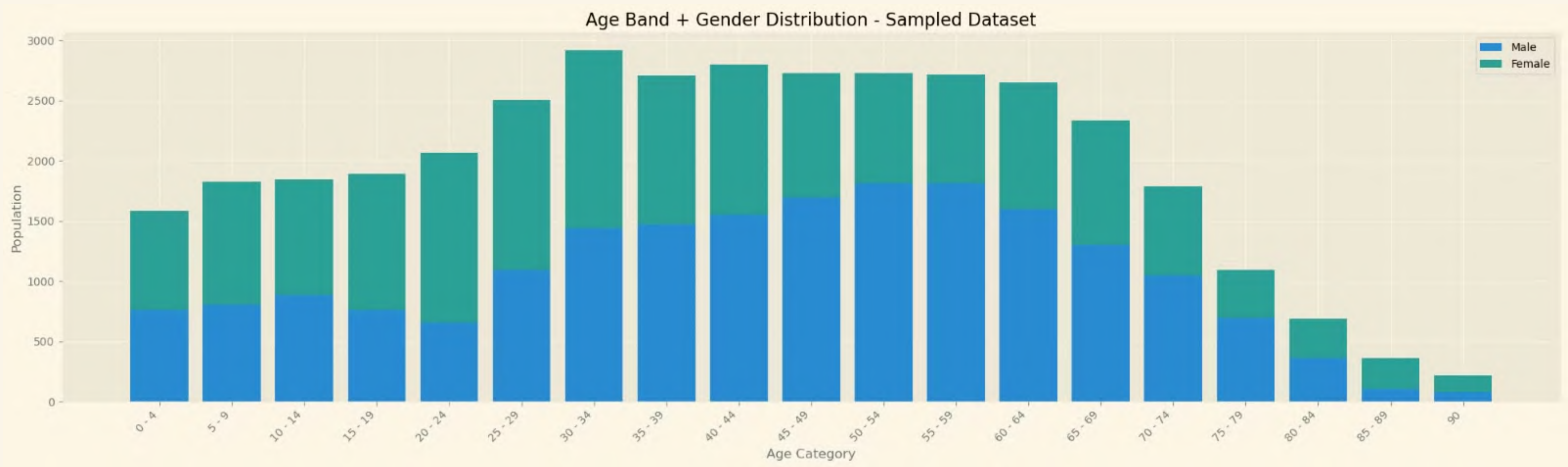
Population Distribution

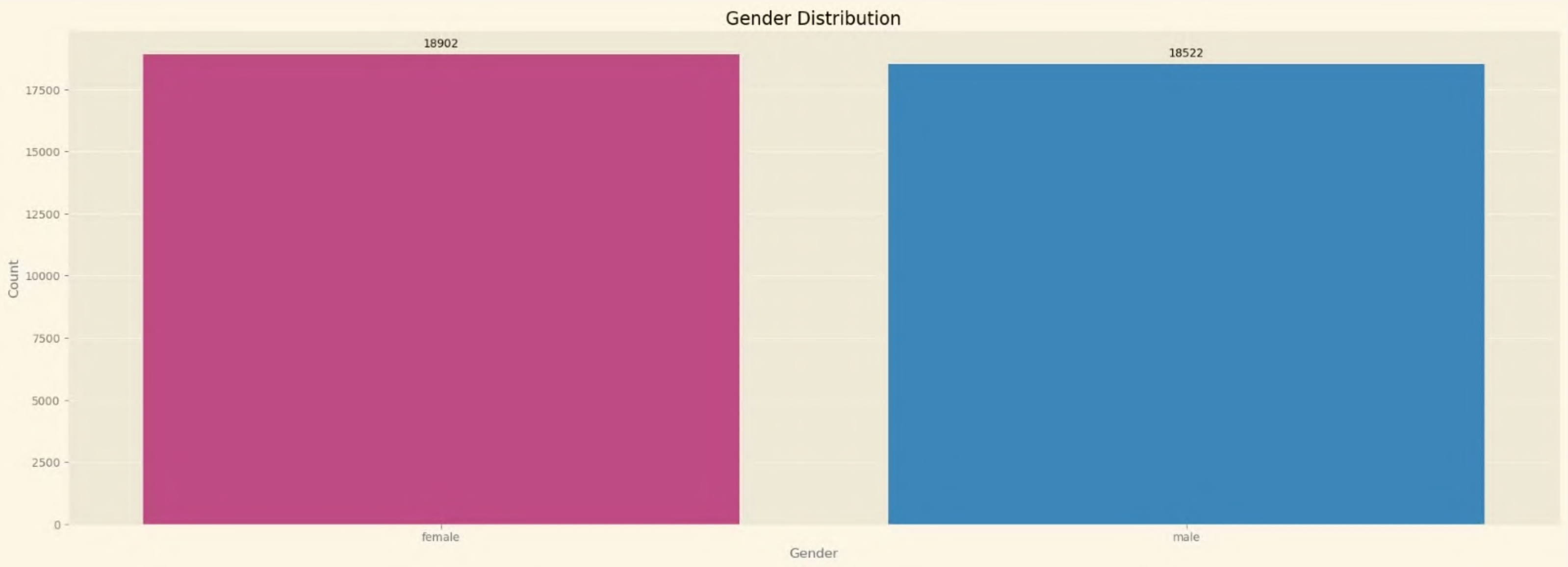


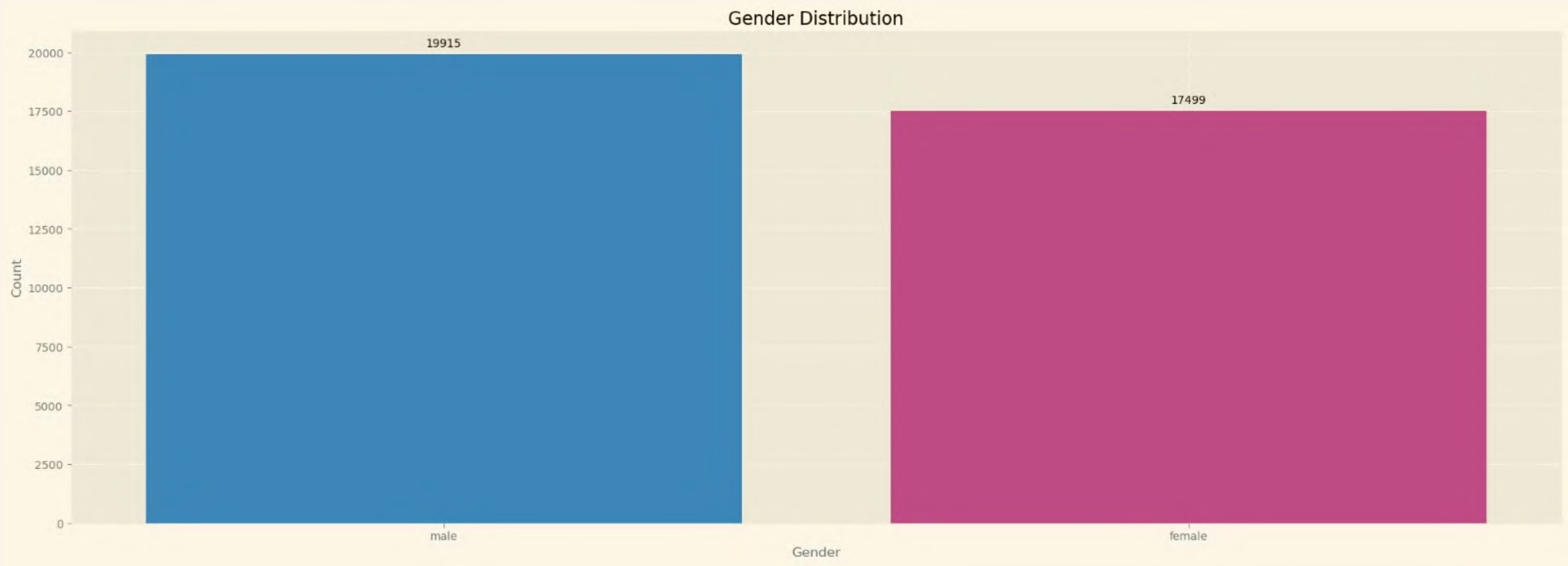
With a compilation of various datasets creating unequal distribution, we chose to mirror Singapore's population distribution (taken from SingStat) to allow for representativeness and reduced bias.

Image Data Sampling





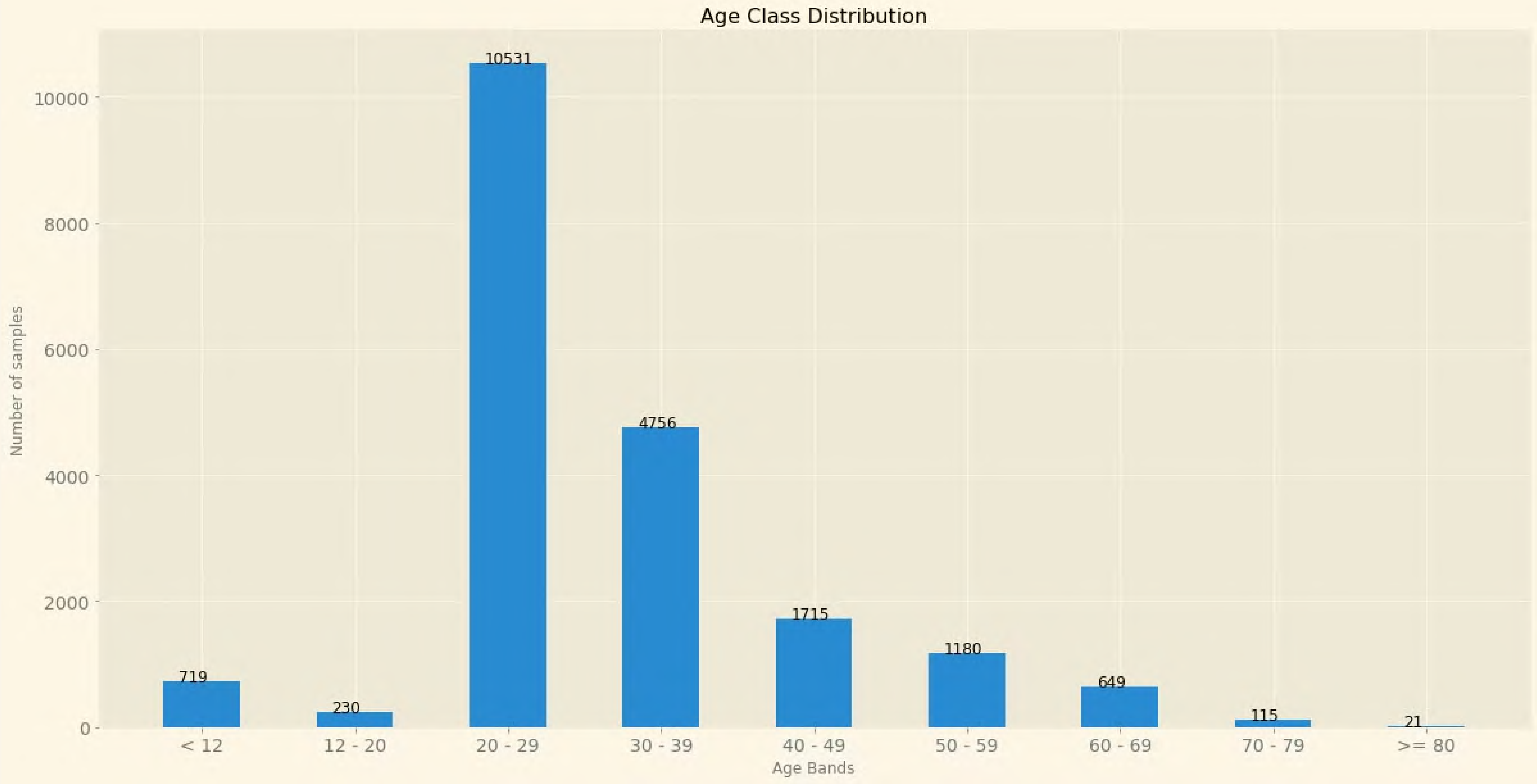




Before Sampling (Raw Distribution)

After Sampling (Sampled Distribution)

Audio Dataset Distribution



Similarly, there was a sincere attempt to distribute the age for our audio training data as well. However, due to the lack of readily available speech datasets with exact age labels and in an effort to keep the size of the dataset relatively adequate, above is the distribution of our final dataset.

Model Exploration

Traditional Machine Learning

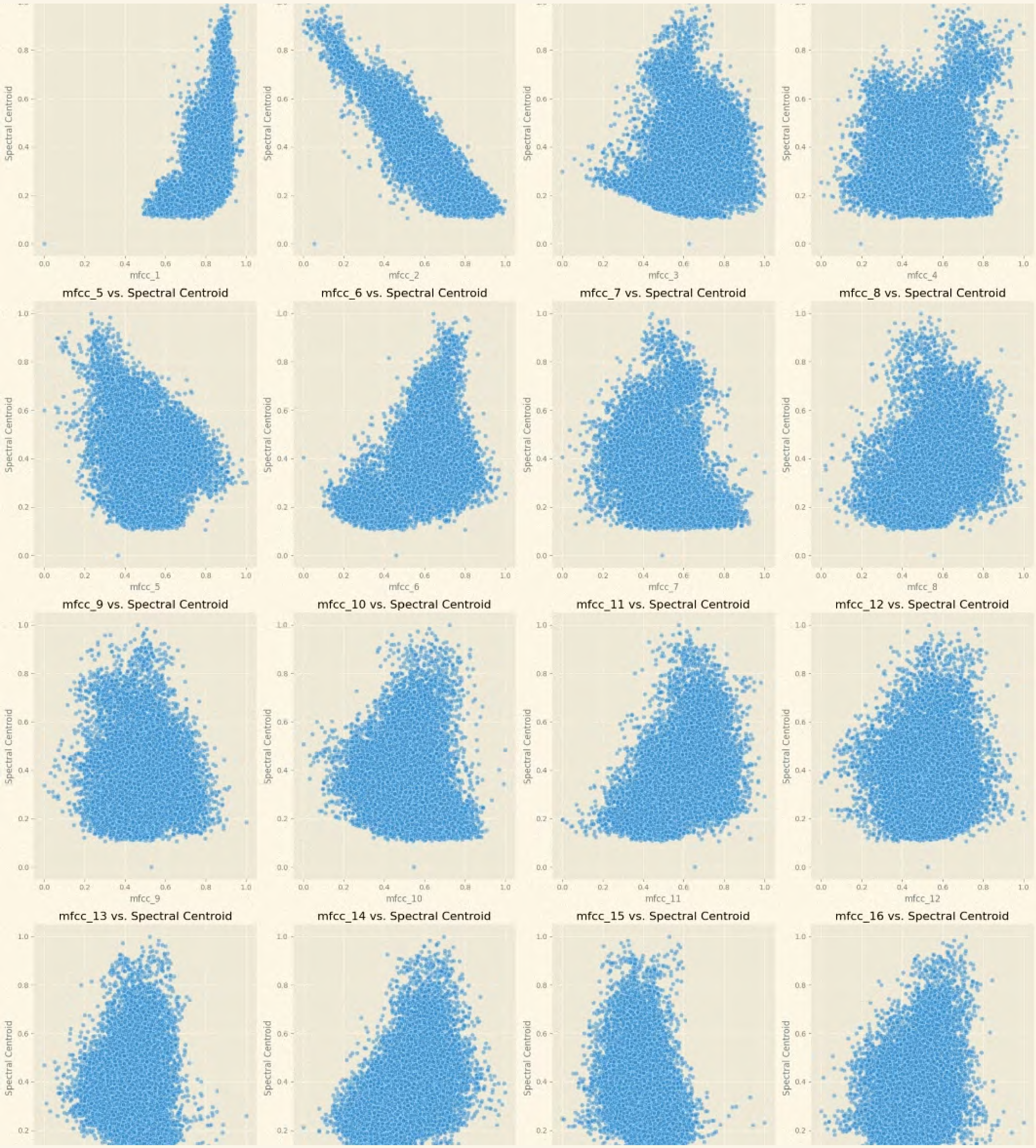
1. MFCC Feature Extraction
2. HOG Feature Extraction
3. Basic Regression Models

Deep Learning

1. Neural Networks
2. Pretrained Models
3. Transfer Learning

Why MSE?

MSE is a versatile and widely accepted loss function that is well-suited for comparing different modeling approaches, especially in regression tasks, where the goal is to minimize the discrepancy between predicted and actual numerical values.



Traditional ML

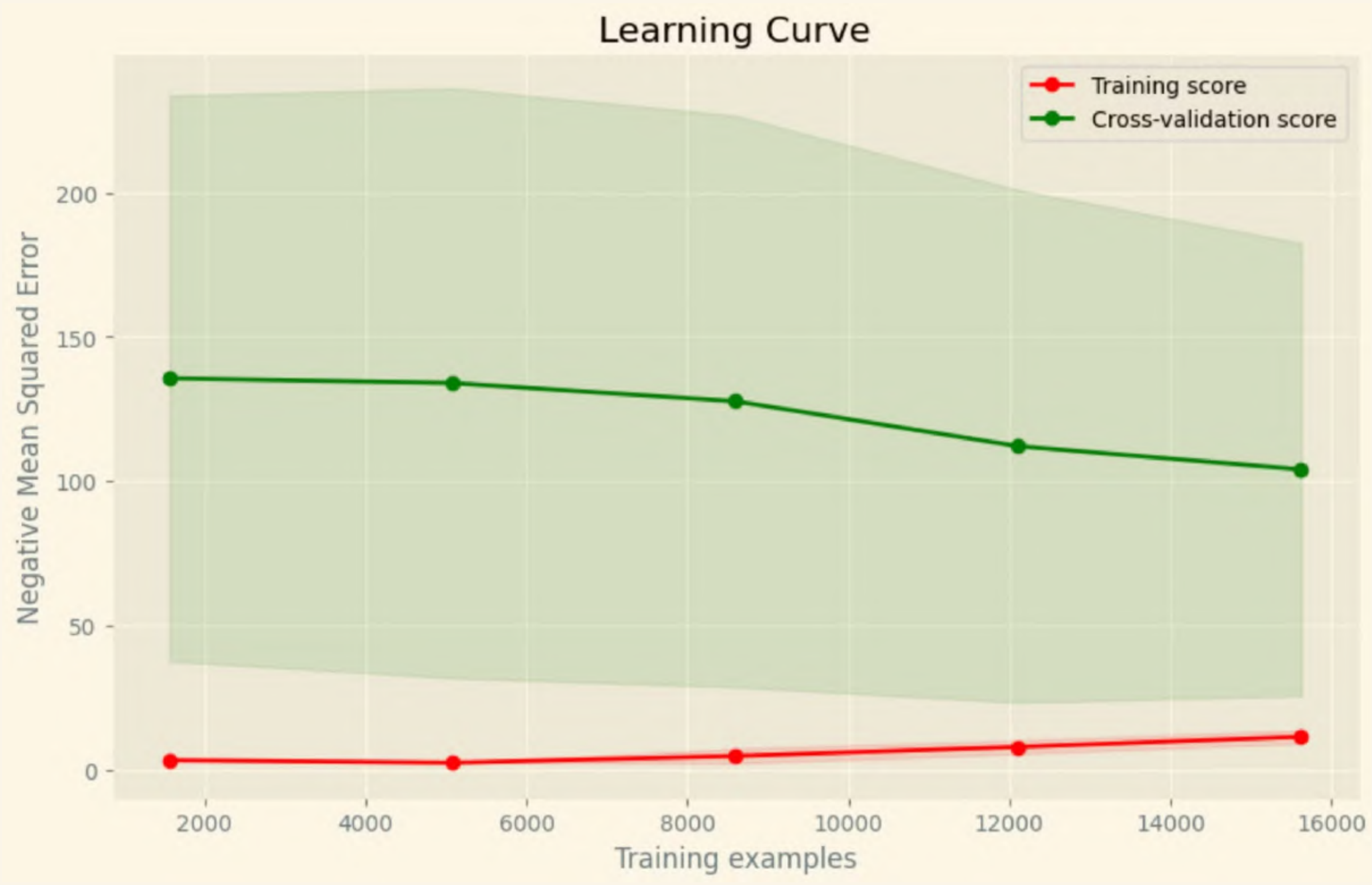
Feature Extraction: MFCC (Audio)

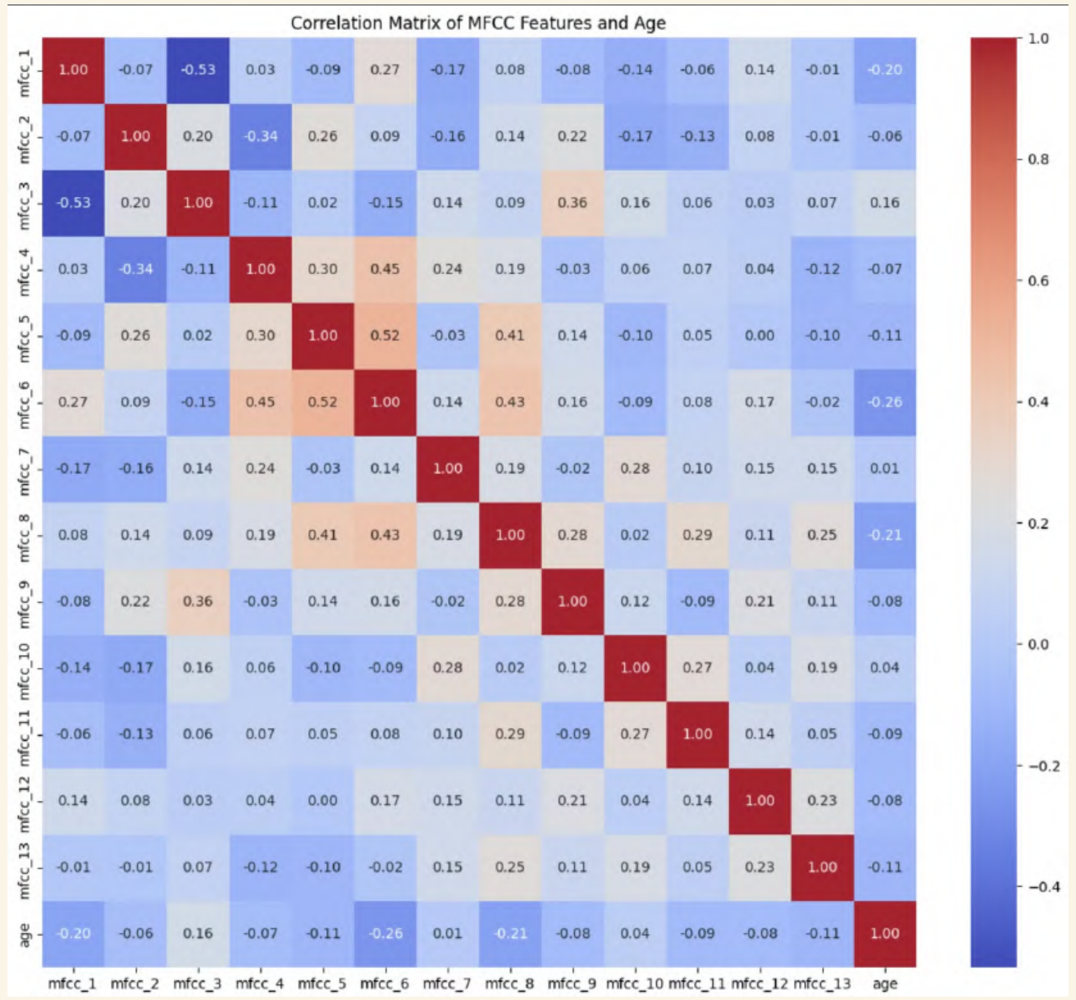
- 20 MFCC
- 7 Spectral Contrast
- 1 Spectral Centroid
- 1 Spectral Bandwidth
- 1 Spectral Rolloff

MFCCs, or Mel Frequency Cepstral Coefficients, are a series of features that represent the shape and characteristics of an audio signal, used extensively in speech and audio processing. They work by mimicking how humans perceive sound, focusing on the most important frequencies and ignoring less significant ones, making them effective for tasks like voice recognition.

Traditional ML (Audio)

Regression Models - Random Forest





Random Forest

Mean Squared Error (MSE): 84.58

Root Mean Squared Error (RMSE): 9.196

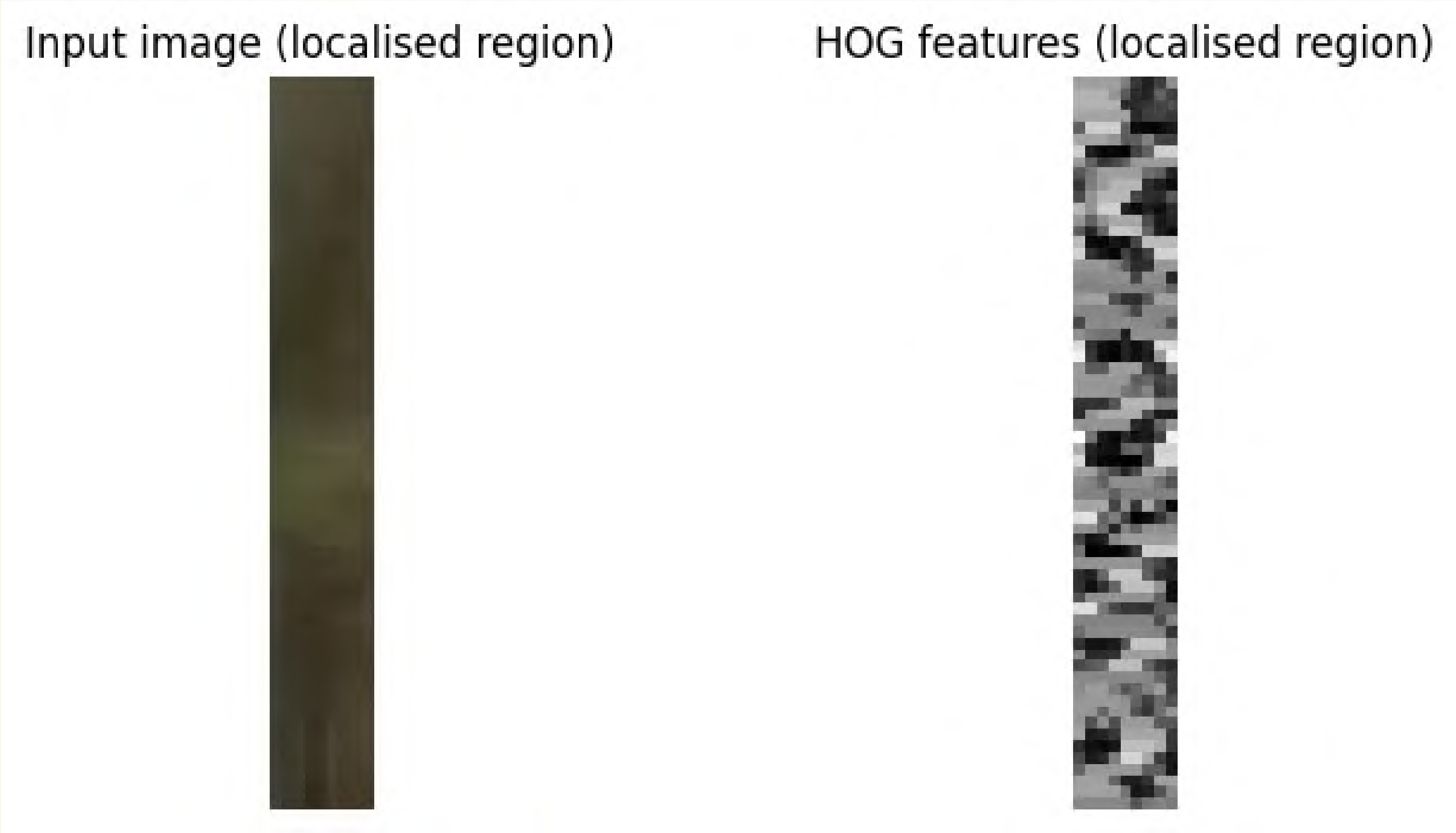
GMM

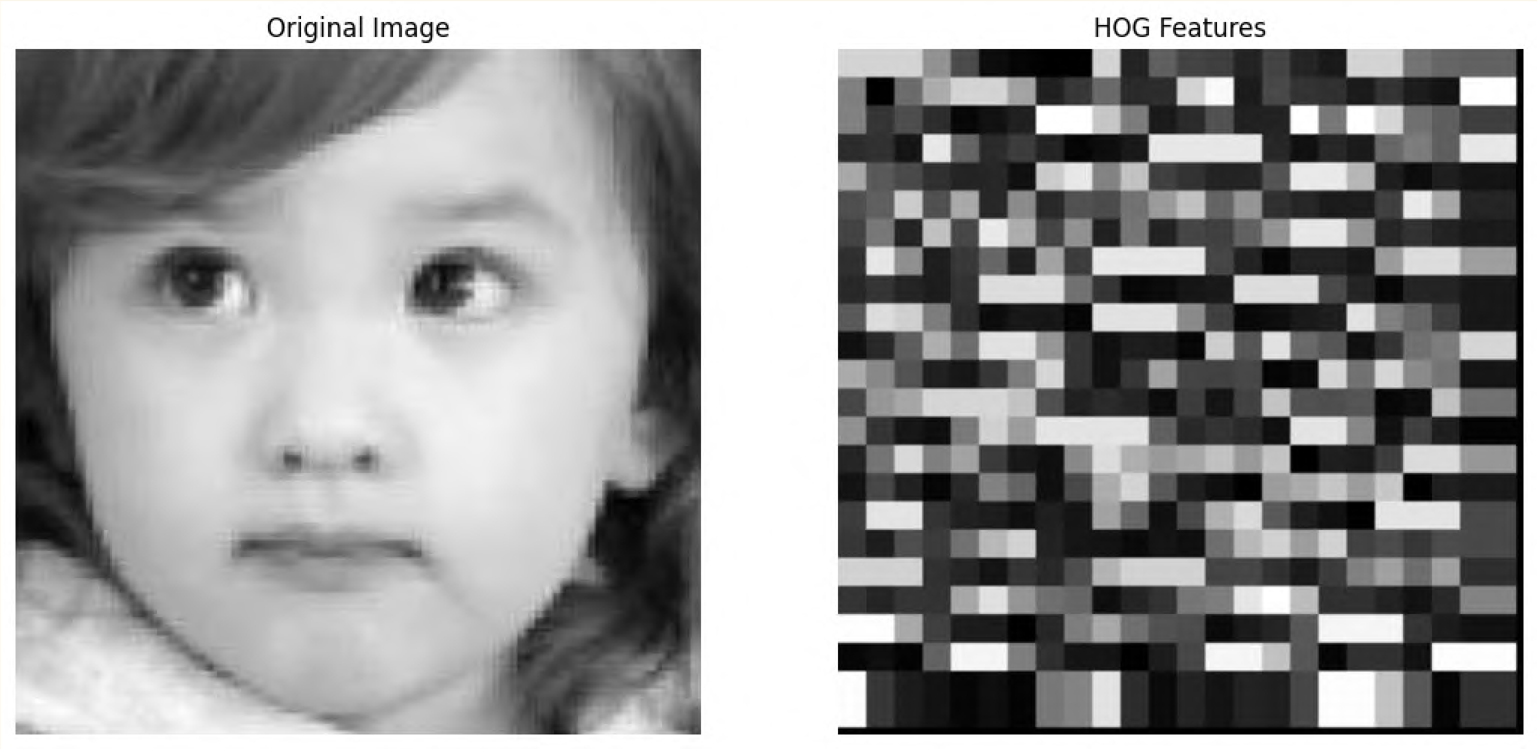
The observed **log likelihood** of training data: -56.98

The observed **log likelihood** of testing data: -48.81

Traditional ML (Image)

Feature Extraction: HOG





- HOG: Feature descriptor in computer vision for object detection, evaluates brightness gradients and compiles them into histograms representing local object appearances and shapes.

- SVR: Type of Support Vector Machine for regression, fits a line with maximum points within a threshold.

- HOG-SVR model: Utilized for tasks like age estimation from photographs, involves extracting HOG features and using SVR to predict age based on these features.

MSE: 358.16297109111883
RMSE: 18.925194083314413

Deep Learning (Audio)

NN for Audio

Approach

Feature extraction was carried out on the audio signals. The following features were extracted:

- Mel-Frequency Cepstral Coefficients (MFCC)
- Spectral Centroid
- Spectral Bandwidth
- Spectral Rolloff
- Spectral Contrast

Results

The maximum accuracy achieved was **64%**

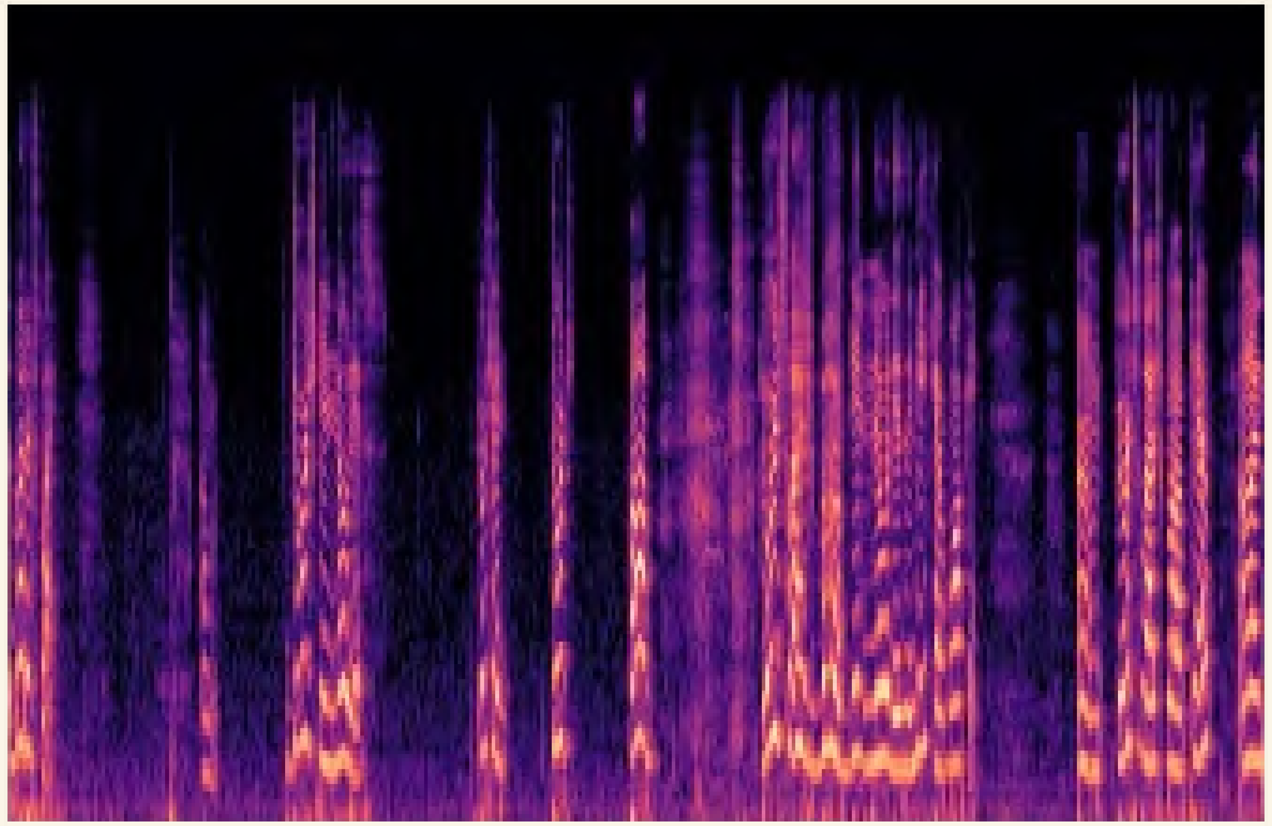
However, the training and validation loss curves were diverging - unable to generalise.

```python
class AgePredictionModel(nn.Module):
    def __init__(self, l1_lambda=0.01, l2_lambda=0.01):
        super(AgePredictionModel, self).__init__()
        self.fc1 = nn.Linear(30, 64)
        self.fc2 = nn.Linear(64, 128)
        self.fc3 = nn.Linear(128, 256)
        self.fc4 = nn.Linear(256, 64)
        self.fc5 = nn.Linear(64, 1)
        self.dropout = nn.Dropout(0.5)
        self.l1_lambda = l1_lambda
        self.l2_lambda = l2_lambda

    def forward(self, x):
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = F.relu(self.fc3(x))
        x = F.relu(self.fc4(x))
        x = self.fc5(x) # No activation, direct regression

        return x
```

Click to collapse the range.

Deep Learning (Audio)

Image Spectrogram



Approach

- Convert each audio signal into a visual representation (spectrogram)
- Use strong image models for age prediction task

VGG16

44%

Train Loss (RMSE):

7.0

Validation Loss (RMSE):

12.0

MobileNetv2

49%

Train Loss (RMSE):

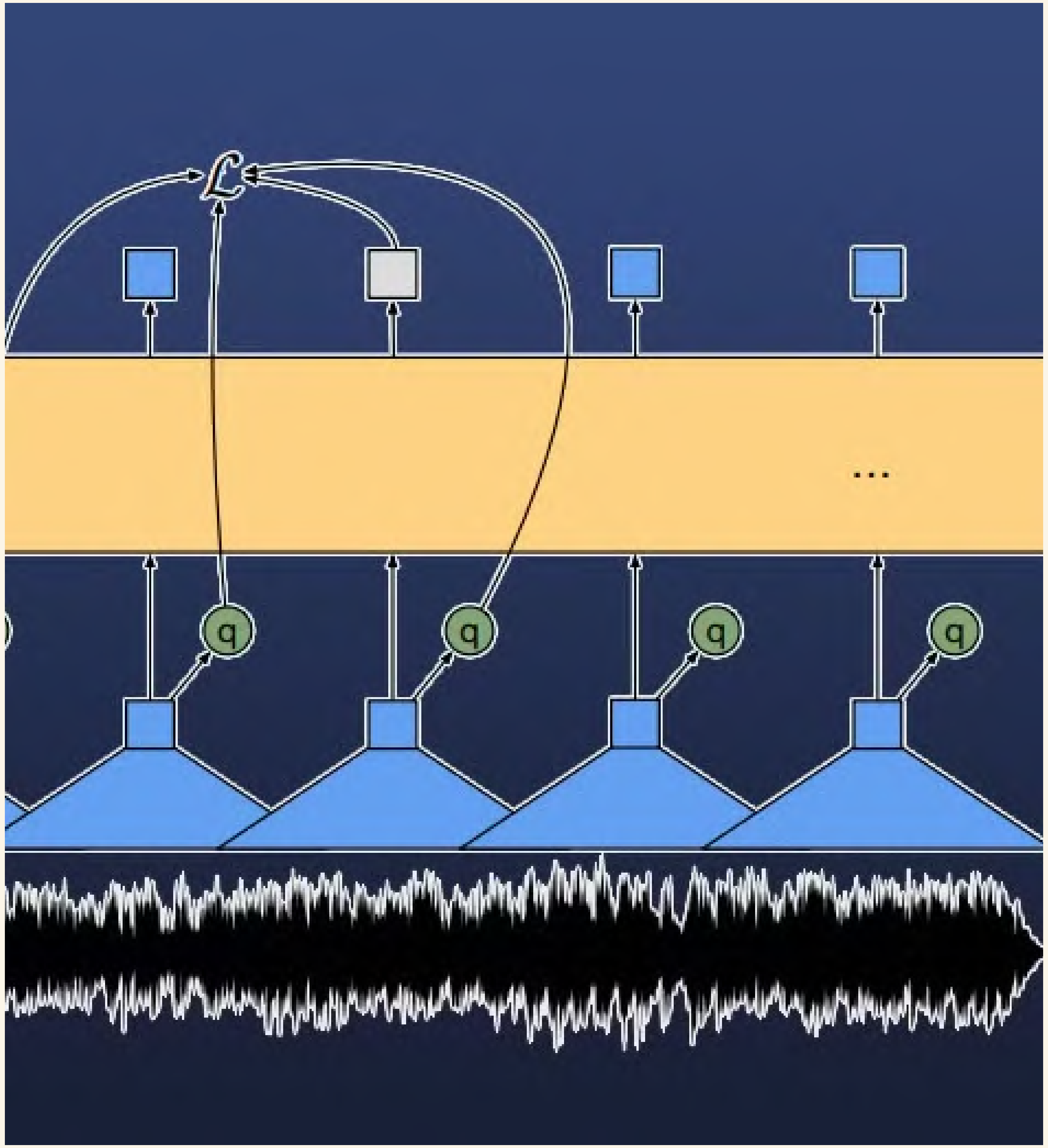
4.0

Validation Loss (RMSE):

10.0

Deep Learning (Audio)

Wav2Vec 2.0



Approach

Adapted wav2vec 2.0 implemented by audEERING GmbH on HuggingFace.

It is a robust 24 layer implementation.

They report metrics: MAE 8.35 years

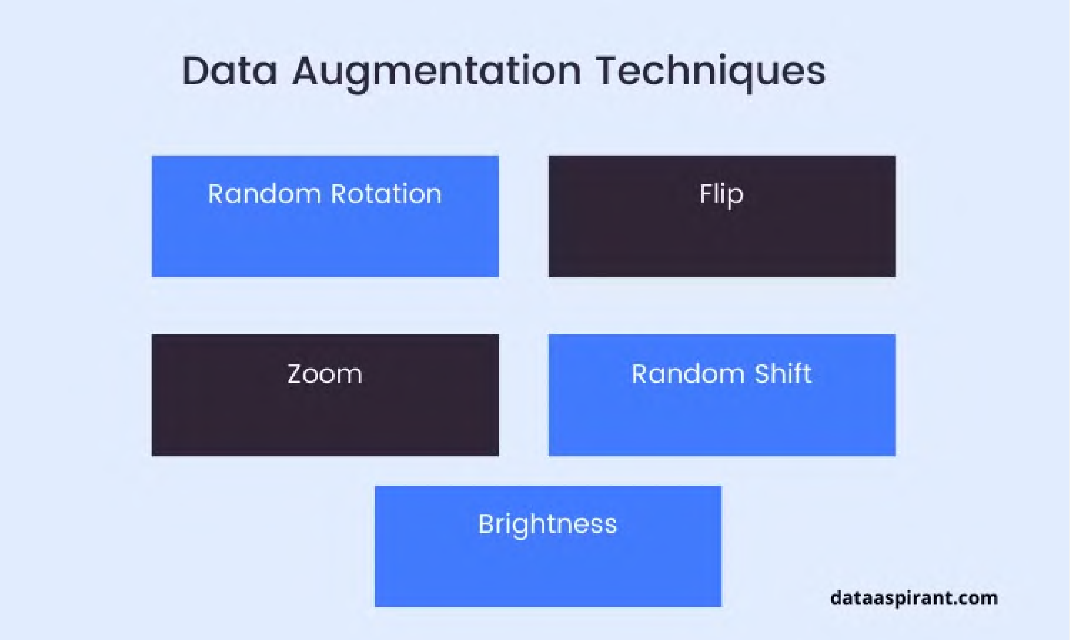
Tested on our dataset of 19,916 samples.

Results

Had an accuracy (+- 6 years) of **55.6%**

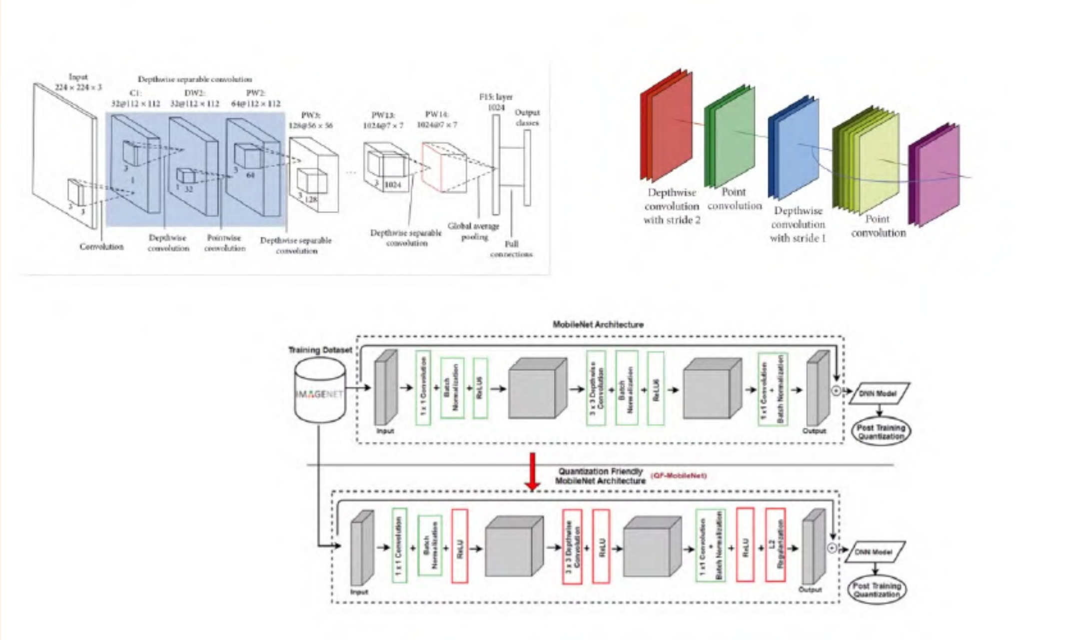
Deep Learning (Images)

Transfer Learning Approach



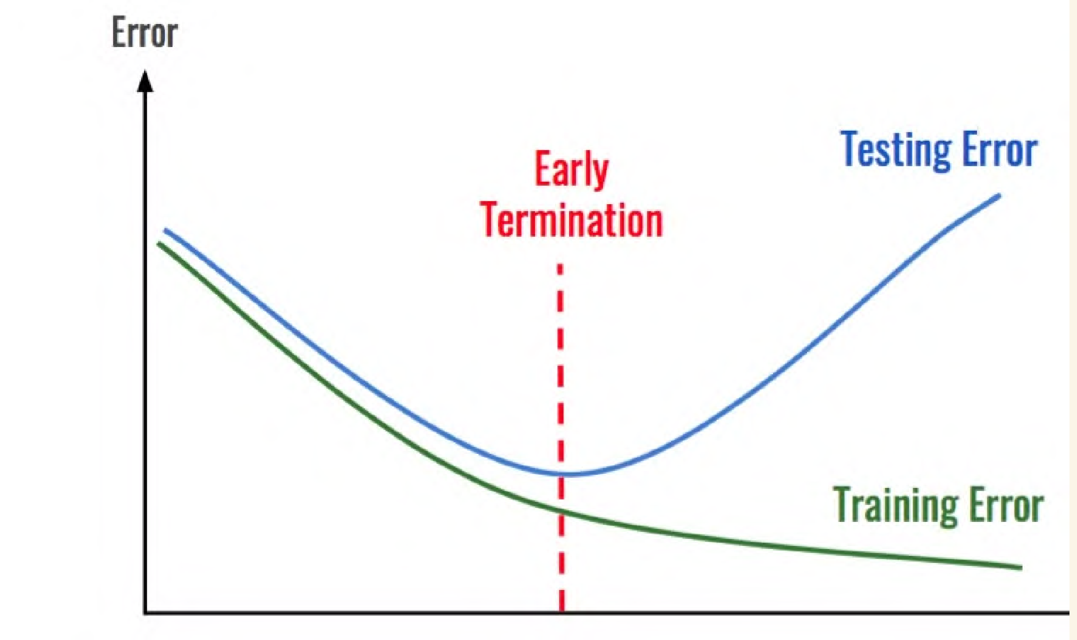
Data Preprocessing and Augmentation

Load images, normalize pixel values, and augment data with techniques like rotation, shifting, and flipping.



Model Architecture setup

Implement **MobileNetV2** with fine-tuning, add custom dense layers for regression, and compile the model with **MSE** and **Adam Optimizer**.



Training and Evaluation

Train the model with **early stopping** and adaptive learning rate reduction, then evaluate its performance on the test dataset by computing the loss.

Deep Learning (Images)

Transfer Learning Results

ResNet50	MobileNetV2 (Most Accurate)	VGG-16	DenseNet	InstructBLIP
31.7%	63.7%	16.6%	18.2%	37.4%
Train Loss: 95.93	Train Loss: 24.73	Train Loss: 515.08	Train Loss: 480.30	Train Loss: 5.82
Validation Loss: 111.94	Validation Loss: 62.27	Validation Loss: 482.03	Validation Loss: 482.11	Validation Loss: 85.00

Fusion Modes (Research)

01

Point of Fusion

1. **Late Fusion**
   - processing each modality separately and then combining the results
2. **Early Fusion**
   - combines raw data from different modalities at the input level (*feature-level* fusion)
3. **Joint Fusion**
   - modalities are processed together throughout the entire pipeline, with the model learning to capture the interactions between them
4. **Common Space Fusion**
   - data from different modalities are mapped into a shared feature space *(cross modal embeddings)*

02

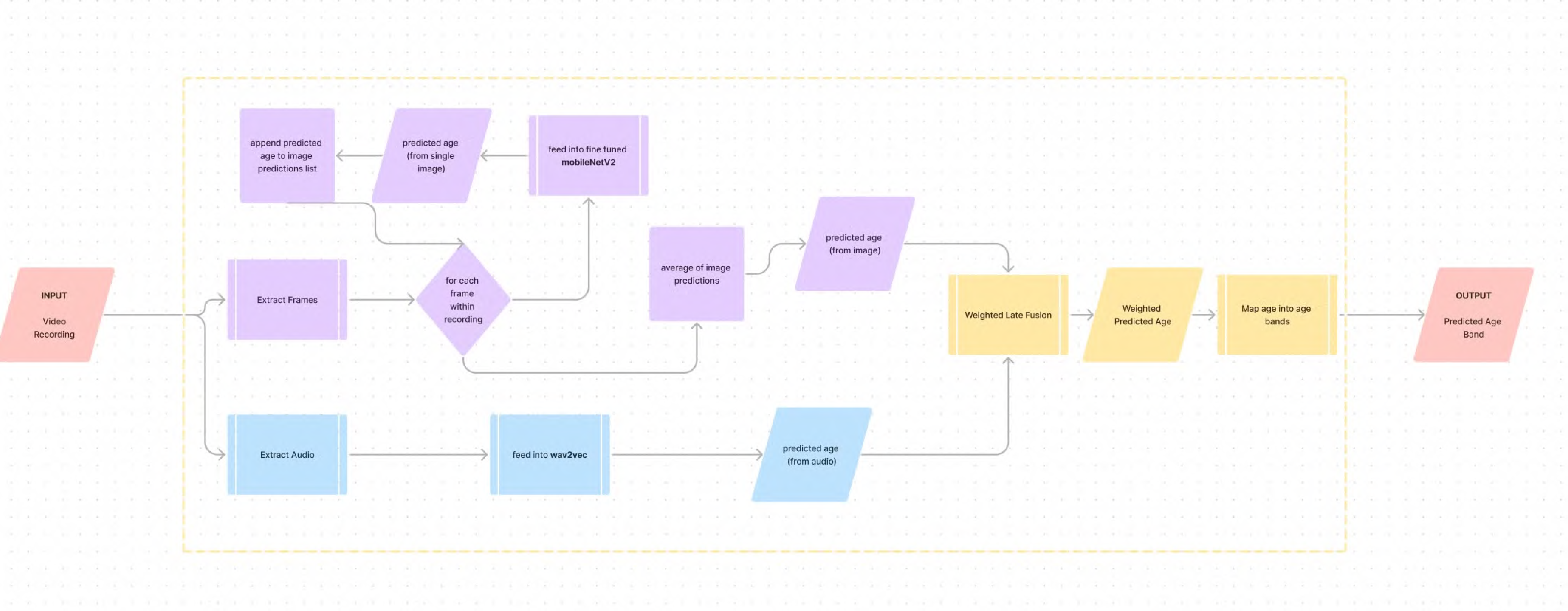
Types of Late Fusion

1. **Summation**
   - features are summed up or concatenated
2. **Averaging**
   - find average output amongst modalities
3. **Voting**
   - feature voting based on modality with highest accuracy
4. **Classifier/Weighted**
   - each classifier produces a probability distribution over the classes, and the final decision is made by combining these distributions

Final Model

Model

1. Architecture
2. Results & Evaluation
3. GUI



Final Model

Weighted Late Fusion, with fine-tuned *MobileNetV2* & *wav2vec2*.

Model Evaluation

Class Imbalance in Audio Dataset

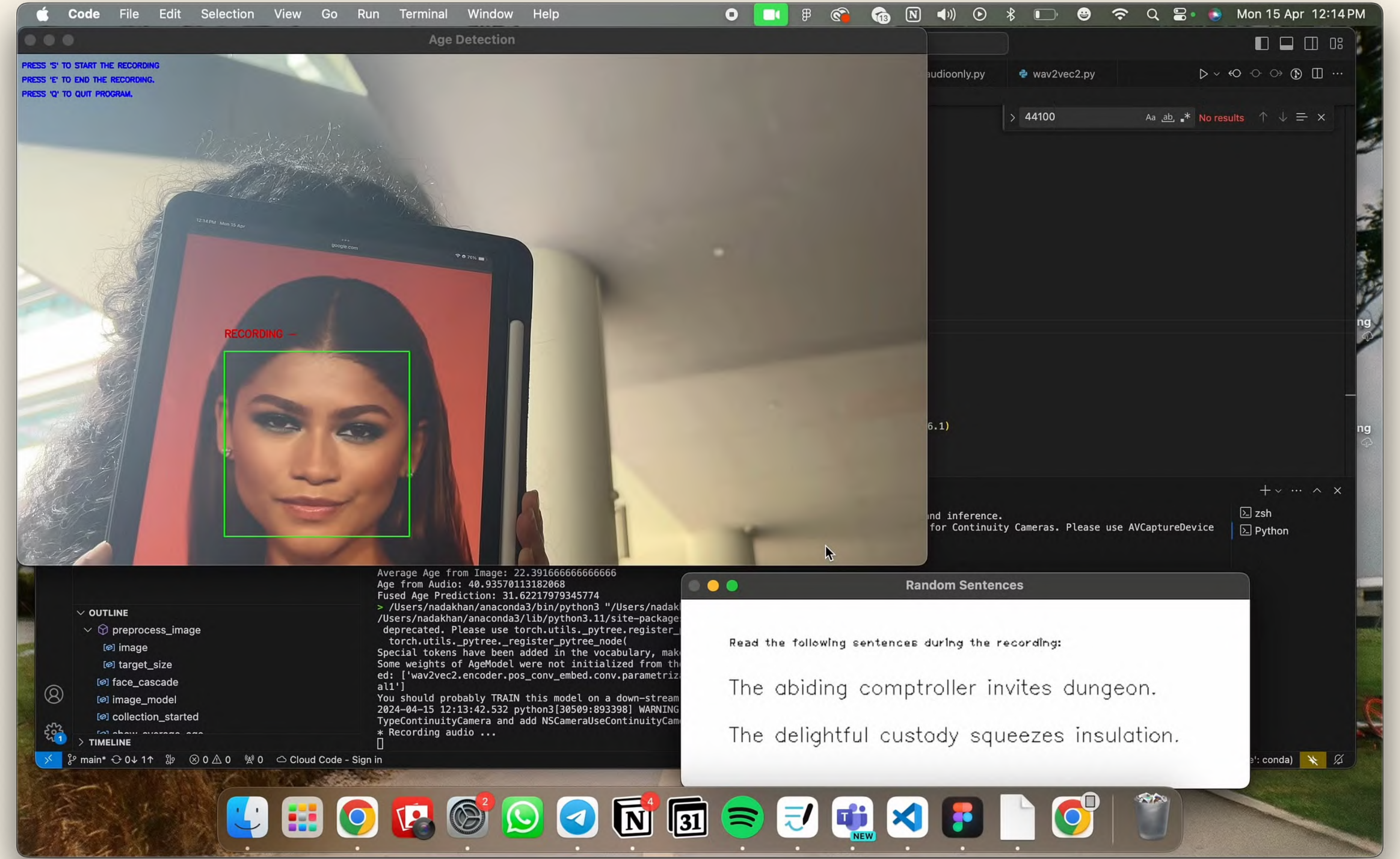
The audio dataset is class-imbalanced. Given more time and resources, this issue can be resolved in the future.

Lack of Datasets with Image, Audio, and Exact Age

Many datasets which are available publicly contain images and audio at most but there are no age labels which are necessary for our project.

Further refining transfer learning on models:
InstructBLIP

Implement techniques to prevent overfitting like regularization and adaptive learning rate reduction.


Code File Edit Selection View Go Run Terminal Window Help
Mon 15 Apr 12:14 PM
Age Detection
PRESS 'S' TO START THE RECORDING
PRESS 'E' TO END THE RECORDING.
PRESS 'Q' TO QUIT PROGRAM.
RECORDING
wav2vec2.py
44100
No results
Average Age from Image: 22.391666666666666
Age from Audio: 40.93570113182068
Fused Age Prediction: 31.62217979345774
Special tokens have been added in the vocabulary,
Some weights of AgeModel were not initialized from
You should probably TRAIN this model on a down-stream
* Recording audio ...
OUTLINE
preprocess_image
image
target_size
face_cascade
image_model
collection_started
TIMELINE
main*
Cloud Code - Sign in
zsh
Python
Random Sentences
Read the following sentences during the recording:
The abiding comptroller invites dungeon.
The delightful custody squeezes insulation.

GUI + DEMO

Thank You!