Age Prediction from Multi-modal (Face and Speech) Input

50.038 Computational Data Science: Final Report

Group 6

Name	Student ID
Chavi Mangla	1005803
Nada Khan	1006212
Radhi Priya Janakiraman	1006387
Rukmini Manojkumar	1005386
Shwetha Iyer	1006308

https://github.com/ssiyer4/CDS_Proj

Contents

1 Introduction **2**

2 Dataset **2**
- 2.1 Collection . . . 2
- 2.2 Sampling . . . 3
- 2.3 Data Pre-processing . . . 4
  - 2.3.1 Feature Extraction for Image . . . 4
  - 2.3.2 Feature Extraction for Audio . . . 5

3 Model Exploration **5**
- 3.1 Traditional Machine Learning . . . 5
  - 3.1.1 Image Models . . . 5
  - 3.1.2 Audio Models . . . 5
- 3.2 Deep Learning . . . 6
  - 3.2.1 Image Models . . . 6
  - 3.2.2 Audio Models . . . 8
- 3.3 Fusion Modes . . . 10

4 Final Model **11**
- 4.1 User Interface . . . 11

5 Results **11**

6 Discussion **12**

7 Future Improvements **12**
- 7.1 Refining Transfer Learning on Models . . . 12
- 7.2 Class Imbalance in Audio Data . . . 12

8 References **12**

1 Introduction

In today's digital age, the ability to extract meaningful insights from multimedia data has become increasingly valuable across various domains. One such area of interest is age prediction where advancements in technology enable us to infer the age of individuals based on visual and auditory cues. The necessity for age prediction using multimodal inputs like images and audio arises from wide-ranging applications and implications across different sectors. From target advertising and personalized content recommendations to healthcare assessment and biometric authentication, the ability to accurately predict age from multimedia data holds immense potential for enhancing user experiences and informing decision-making processes. In the realm of online safety, the pressing issue revolves around implementing reliable age verification mechanisms through age verification to prevent underage users from accessing potentially harmful or inappropriate content, thus fostering a safer digital environment for individuals of varying age groups. This project delves into the development and exploration of age prediction models leveraging on a multimodal approach that relies on video input.

2 Dataset

2.1 Collection

Obtaining a diverse and representative dataset for training our multi-modal age prediction model posed several challenges. Given the complexity of acquiring clean and varied video data, we adopted a hybrid approach by compiling samples from both image and audio datasets as separate modalities. This decision was made to ensure enough data for training while maintaining diversity across different modalities.

Our dataset comprises samples sourced from various publicly available datasets, which can be summarized in the table below:

Audio Speech Datasets	
CREMA-D *(CheyneyComputerScience, n.d.)*	An emotional multimodal actor dataset of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 73 coming from a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic and unspecified).
english_children *(James et al., 2016)*	The dataset contains audio recordings (lossless WAV) of 11 young children (age M=4.9 years old; 5 females, 6 males).
Eugene Children's Story Corpus (ECSC) *(Kallay, J., & Redford, M. A., 2020)*	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.
SpeechAccent *(Speech Accent Archive, 2017)*	This dataset contains 2140 speech samples, each from a different talker reading the same reading passage. Talkers come from 177 countries and have 214 different native languages. Each talker is speaking in English.

Table 1: *Details of speech audio datasets*

Image Datasets	
Dataset	Description
UTKFace *(UTKFace, n.d.)*	A large-scale face dataset with long age span. The dataset consists of over 20,000 face images with annotations of age, gender and ethnicity. The images cover large variations in poses, facial expression, resolution etc.
All-Age-Faces *(JingchunCheng, n.d.)*	Contains 13,322 face images (mostly Asian) distributed across all ages (from 2-80) including 7381 females and 5941 males.

Table 2: *Details of face image datasets*

2.2 Sampling

The final dataset consists of 57,330 samples - 19,916 audio samples and 37,414 image samples.
To address class imbalances from the various datasets and to maintain diversity, we ensured that our sampled data mirrored the distribution of population statistics – this is done in the context of Singapore, with data from SingStat *(Population and Population Structure, n.d.)*.

Here, we examine the population distribution of Singapore:

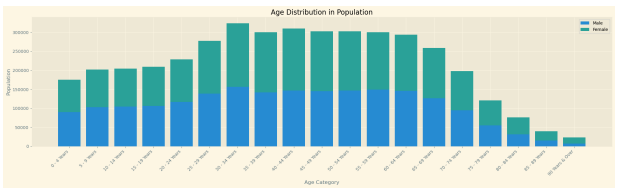

Figure 1: *Population distribution of Singapore*

However, the initial collected facial image dataset was not well-varied:

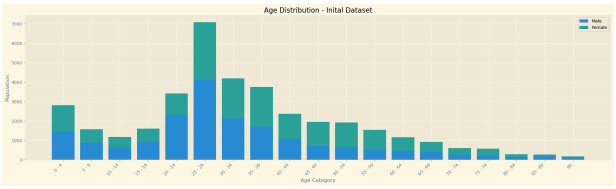

Figure 2: *Initial age distribution of facial image dataset (top right of the image)*

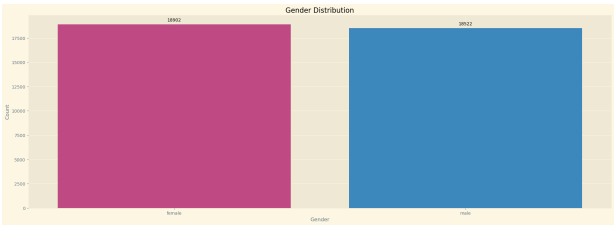

Figure 3: *Initial gender distribution of image data*

According to the population distribution in Figure 1, our dataset of facial images was re-sampled to mirror the population distribution.

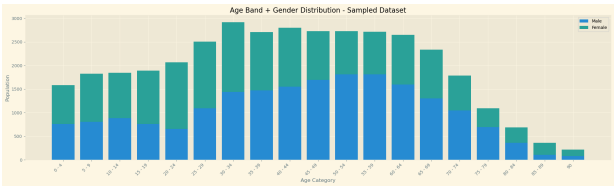

Figure 4: *Age distribution of re-sampled facial image dataset (top right of the image)*

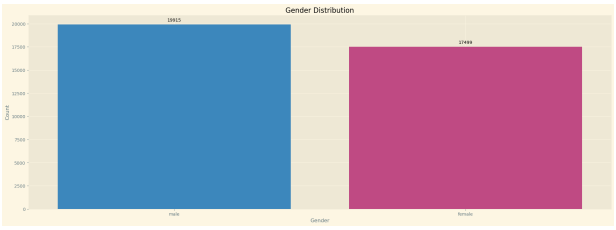

Figure 5: *Gender distribution of image data after re-sampling*

Our samples were banded for the sake of classifying and mirroring the banded age distributions as extracted from SingStat *(Population and Population Structure, n.d.)*, however as we want to do an age regression, our sampled input distribution can be visualized by specific ages as below:

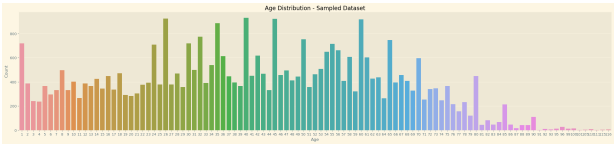

Figure 6: *Age distribution by individual age*

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 adequate, this is the distribution of our final dataset:

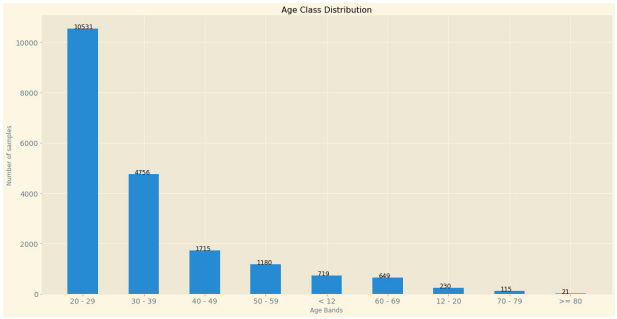

Figure 7: *Distribution of ages of audio dataset*

To evaluate the performance of our age prediction model effectively, we divided the dataset into training, validation, and testing sets. Stratified sampling based on age was employed to ensure that each set maintains a representative distribution. The training set comprises 80% of the data, while the validation and testing sets consist of 10% each.

2.3 Data Pre-processing

2.3.1 Feature Extraction for Image

We opted to apply a series of filters, including grayscale conversion, Gaussian blurring, Gabor filtering, and Sobel + Canny edge detection, to explore potential features relevant to age prediction within facial images.

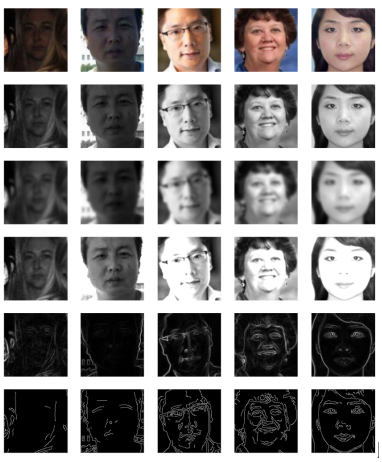

Figure 8: *Features extracted from image data*

After applying various filters to a sample of five images, it became apparent that each filter highlighted different facial features. Particularly, while the canny edges filter did capture distinctions within age groups by resembling wrinkle lines, it was noted that this method might not be the most reliable due to inconsistencies stemming from variations in image quality and contrast. For instance, one of the images displayed minimal lines (most-left sample), leading to challenges in accurately extracting features.

Considering this, we explored Histogram of Oriented Gradients (HOG) as a means of image feature extraction, which may be a more reliable basis of features for our model. HOG works by capturing the local gradient information through quantifying the distribution of gradient orientations within localized regions– HOG features can be extracted from images using the scikit-image library, and visualized as below:

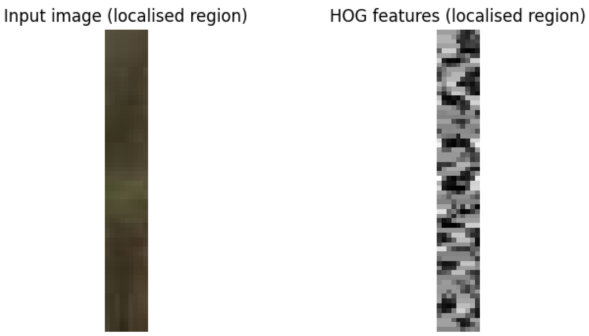

Figure 9: *Feature extraction from images - HOG*

HOG is usually done on localized regions as shown in Figure 9. However for a simple example, a full image is passed here to visualize how a full image could be represented:

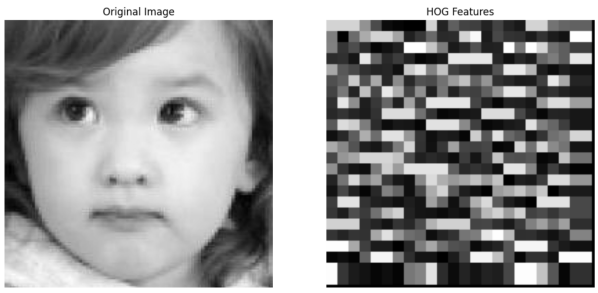

Figure 10: *HOG representation of a full image*

2.3.2 Feature Extraction for Audio

For the task of age prediction, feature extraction plays a pivotal role in transforming raw audio data into a comprehensible format that deep learning models can efficiently process. This process involves converting raw audio signals into a set of representative features, capturing the essential characteristics of the audio signal relevant to the task of age prediction. For the task at hand, these are the features extracted (using the Python library librosa):

i. **Mel-Frequency Cepstral Coefficients (MFCCs):** represent the energy distribution across different frequency bands, according to the Mel-scale.

ii. **Spectral Centroid:** indicates where the "center of mass" of the spectrum is located. In other words, it represents the weighted mean of the frequencies present in the signal.

iii. **Spectral Bandwidth:** measures the width of the frequency range in which most of the signal's energy is concentrated.

iv. **Spectral Rolloff:** is a measure of the frequency below which a certain percentage (typically 85-95%) of the total spectral energy is contained.

v. **Spectral Contrast:** measures the difference in magnitude between peaks and valleys in the spectrum of an audio signal.

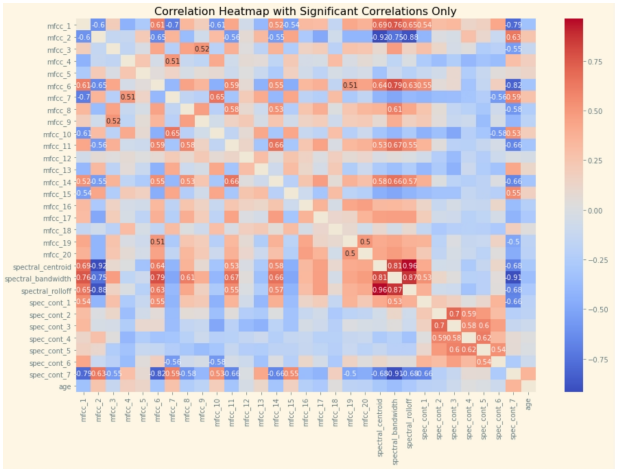

Figure 11: *Correlation heatmap of audio features*

3 Model Exploration

In this section, we explore various approaches to create a robust age prediction model, ranging from traditional machine learning methods to more modern transfer learning techniques. Each approach offers distinct advantages and challenges, which we analyze in the context of our multimodal age prediction task.

3.1 Traditional Machine Learning

3.1.1 Image Models

With these HOG features, we trained a support vector regression (SVR) model that had the following results:

Metric	Value
Mean Absolute Error (MAE)	15.37
Mean Squared Error (MSE)	358.16
Root Mean Squared Error (RMSE)	18.96

Table 3: *Results of Image SVR model*

3.1.2 Audio Models

In the training of models using the audio data, we initially employed two distinct regression approaches to predict age based on vocal characteristics: a Gaussian Mixture Model (GMM) and a Random Forest classifier, each utilizing 13 Mel Frequency Cepstral Coefficients (MFCCs) as features.

i. **Gaussian Mixture Model (GMM)** The GMM approach, which models the distribution of the MFCCs using a mixture of Gaussian distributions, was anticipated to effectively capture the nuances in audio data attributable to different age groups. A GMM with five pre-defined components was initialized and trained on the compiled feature set. Evaluation of the GMM model's performance relied on the log-likelihood metric, offering insights into the model's fit to the training data and its ability to generalize to unseen testing data. As seen in the results

below, its performance was sub-optimal.

Log-likelihood Metric	Value
Training	-56.98
Testing	-48.81

Table 4: *Results of GMM model*

These findings underscore the necessity of increasing the complexity of the models or potentially integrating more discriminative features to improve the accuracy of age prediction based on vocal attributes. This could involve exploring additional or alternative feature sets that capture more detailed aspects of the voice or adjusting model parameters to enhance their learning capacity.

Subsequently, Principal Component Analysis (PCA) was performed, and a scatter plot of the data reduced to two principal components derived from MFCC features was created. The PCA graph indicates that the features, when represented in the reduced PCA space, do not demonstrate clear boundaries between different age groups. This lack of separation suggests that the GMM, in conjunction with diagonal covariance, may not possess sufficient complexity to adequately model the data distribution, potentially leading to overfitting.

Figure 12: *Principal component analysis on MFCC features*

ii. **Random Forest Classifier**

The Random Forest classifier, known for its robustness and ability to handle overfitting through ensemble learning, was implemented alongside the Gaussian Mixture Model (GMM) to assess its effectiveness.

Metric	Value
Mean Squared Error (MSE)	84.58
R-sqaured value	0.35

Table 5: *Results of random forests model*

To gain insight into the model's behavior, a scatter plot and a learning curve were plotted.

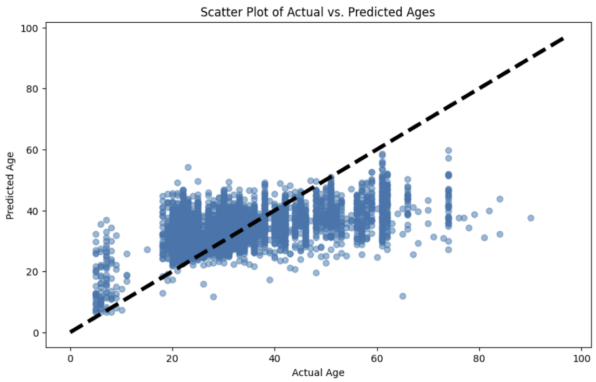

Figure 13: *Scatter plot result of random forests result*

The scatter plot visualization of actual vs. predicted ages shows that the model's predictions are not perfectly aligned with the ideal scenario. The plot exhibits broad dispersion, reflecting variability in the model's age prediction accuracy across different age groups. Predictions are denser around the 20 - 60 age range, hinting that the model performs better for middle-aged individuals. The spread of predictions for a given actual age indicates inconsistencies in the model, as seen in the wide variance along the y-axis.

3.2 Deep Learning

Aside from exploring traditional methods of machine learning that involved simple feature extraction and regression analysis, deep-learning methods were also explored.

3.2.1 Image Models

I. Convolutional Neural Network for Image

A CNN for age regrssion was developed. The model uses the Mean Squared Error (MSE) as its loss function during training to improve accuracy by closing the difference between predicted and true ages. The images were pre-processed to resize them to 128x128 pixels.

The CNN implemented consists of two convolutional layers followed by max-pooling for downsampling, along with two fully connected layers for regression-based age prediction.

The model leverages Rectified Linear Unit (ReLU) activation functions and dropout regularization to introduce non-linearity and prevent overfitting, respectively, while employing Mean Squared Error (MSE) as the loss function for training.

The CNN model initially demonstrated promise; the results are detailed below:

Metric	Value
Mean Squared Error (MSE)	85.00
Root Mean Squared Error (MSE)	9.22
Training Loss	5.90
Validation Loss	88.59

Table 6: *Results of CNN for image*

The model performs poorly on unseen data, which can be observed from the validation loss being 88.59 and the training loss being 5.90. Despite its suitability for regression-focused tasks like age prediction, evidenced by the results in Table 6, the observed overfitting prompted a decision to forego its further utilization.

II. Transfer Learning

There has been ongoing research on exploring different ways to predict ages from images using pre-trained models. The models described below, like ResNet50, VGG-16, MobileNetV2, DenseNet, CNN and CLIP, offer different ways to tackle the same problem. This section will describe how each model was trained and how well they predict ages.

Model	Training Loss	Validation Loss	Accuracy
ResNet50	95.93	111.94	31.70%
VGG-16	515.08	482.00	16.60%
MNetv2	24.73	62.27	63.70%
DenseNet	480.30	482.11	18.20%
CLIP	449.02	471.52	22.41%

Table 7: *Results of various models for image*

i. **ResNet50**

The ResNet-50 model is trained using a custom dataset class that loads images, applies various transformations for data augmentation during training, and minimal transformations for validation and testing. The training set undergoes transformations like resizing, random horizontal flips, rotations, and color jittering to enhance the model's ability to generalize to unseen data. These images are then normalized before being fed into the ResNet-50 architecture. The model's last fully connected layer is modified for age prediction, and the entire network is trained using cross-entropy loss and the Adam optimizer.

ii. **VGG-16**

Initially, a pre-trained VGG-16 model without its top (fully connected) layers is loaded, with the intention of leveraging its learned features for the task at hand. The loaded model's layers are then frozen to prevent further training of these pre-existing weights. Next, additional layers are appended to the model to enable age prediction. This includes a flattening layer followed by a densely connected layer with 512 units and ReLU activation, along with dropout regularization to mitigate overfitting. Finally, a single neuron output layer with linear activation is added to predict the age. To train the model, data generators with augmentation are employed. These generators prepro-

cess images, applying various transformations such as rotation, shift, shear, zoom, and horizontal flip, augmenting the dataset to improve model generalization. The training process utilizes early stopping as a regularization technique to prevent overfitting.

iii. **MobileNetv2**

The MobileNetv2 (abbreviated as MNetv2 in Table 7) model is trained for age prediction involving processes such as data pre-processing, augmentation, model customization and training. The image data generators were employed to standardize the image data, ensuring consistency in size and pixel values. Data augmentation such as rotation, shifting, shearing, zooming and flipping are applied to diversify the training data and enhance model generalisation. Fine-tuning of the MobileNetV2 base model involves unfreezing selected layers for customization with additional dense layers added for age prediction. Early stopping was also implemented to prevent overfitting.

iv. **DenseNet**

The DenseNet-121 model, initially trained on ImageNet, has been adapted from classifying discrete categories to estimating continuous age values, leveraging its deep connectivity for effective age prediction from images. Originally designed to output probabilities for 1,000 classes, its final layer was reconfigured to a single output suitable for regression, streamlining the process from class probabilities to direct age estimation. The loss function was transitioned to Mean Squared Error (MSE), ideal for regression as it robustly measures the discrepancies between predicted and actual ages.

To ensure the model's robustness and adapt to various facial features, image augmentation techniques like Random Resized Crop and Random Horizontal Flip were applied, introducing variability and aiding in learning from diverse facial presentations. This is vital for adapting to the heterogeneous nature of human ages. Additionally, normalization using ImageNet's mean and standard deviation ensures the input data aligns with the pre-trained model's expectations, optimizing feature utilization.

v. **CLIP**

The CLIP model, initially trained to understand a broad range of images with textual descriptions, has been adapted for age prediction by replacing its classification head with a regression layer. This change allows the model to estimate ages as a continuous value rather than categorize into classes. The model now uses Mean Squared Error (MSE) for training, which effectively minimizes the error between predicted and actual ages. Image augmentation techniques like Random Resized Crop and Random Horizontal Flip are employed to enhance robustness and adaptability to variations in image alignment and scale. Additionally, the images are normalized using ImageNet's mean and standard deviation to maintain compatibility with pre-trained weights, optimizing the model's performance for age estimation.

3.2.2 Audio Models

I. Neural Network for Audio

The initial deep-learning approach to the age prediction task was to use a neural network with the extracted features mentioned in section 1(b). These thirty input features were fed into a feed-forward neural network with five fully connected layers. Dropout and elastic regularization (L1 + L2

regularization) techniques were applied as well.

The network was trained with PyTorch and an Adam optimiser (weight decay factor = 0.0005). The train-val-test split was 0.8, 0.1, and 0.1 respectively. The loss function used was Root Mean Squared Error (RMSE). Additionally, hyperparameter tuning was carried out in the form of grid search, with the hyperparameters being learning rate, batch size, and number of epochs.

Over many iterations and runs, the highest accuracy achieved was 64% (accuracy is measured as the prediction being within 6 years of the actual age). The final train loss was 6.7 years and the final validation loss was 11.7 years, after 170 epochs with a learning rate of 0.005, these were the best results obtained:

Metric	Value
Accuracy	64.00%
Training Loss	6.7
Validation Loss	11.7

Table 8: *Results of NN for audio*

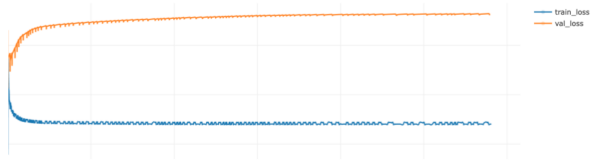

Figure 14: *Training and validation loss curves for NN for audio*

However, as seen in Figure 13, the training and validation curves diverge before plateauing. This suggests that the network is not able to generalize well, and unable to recognize the complex relationships between the data points. Thus, this approach was abandoned.

II. Transfer Learning for Audio

The second approach for the task of age prediction from voice was to convert the audio signals into visual representations in the form of spectrograms. These images could then be passed through strong vision models. The models experimented with for this approach were VGG-16 and MobileNetv2.

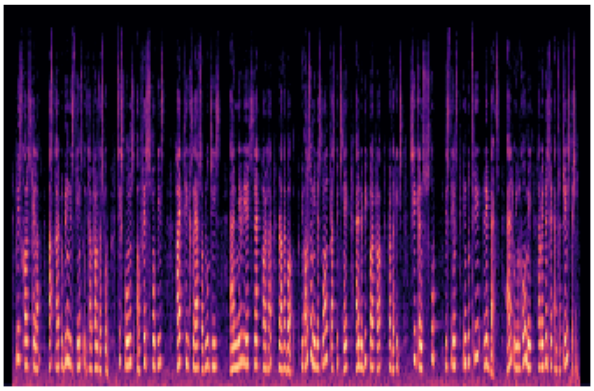

Figure 15: *Sample spectrogram*

For VGG-16, the model's architecture retained the convolutional base of VGG16, including the feature extraction layers and the average pooling layer. The classifier section was modified to predict age, consisting of a linear layer that reduces dimensionality to 256, followed by ReLU activation and dropout for regularization, culminating in a final linear layer that outputs a scalar prediction of age.

The second experiment was built on the MobileNetV2 (abbreviated as MNetv2 in Table 9) architecture known for its lightweight yet powerful capabilities. This model utilized a pre-trained MobileNetv2 version to harness learned features from extensive datasets, optimizing performance without intensive training. The classifier of the MobileNetV2 was modified to accommodate a specific task by replacing the original classifier layer with a new linear layer that maps 1280 input features to a single output, making it ideal for age prediction.

For both experiments, the models were trained on PyTorch with an Adam optimizer. The loss function used was Root Mean Square Error (RMSE). The table below details the results of the experiments:

Model	Training Loss	Validation Loss	Accuracy
VGG-16	7.0	12.0	44%
MNetv2	4.0	10.0	49%

Table 9: *Results of various models for image*

III. Wav2Vec 2.0

The third and final approach for this task was leveraging the wav2vec2-large-robust-24-ft-age-gender developed by audeering and hosted on HuggingFace. This model exploits the capabilities of Wav2Vec2, a pre-trained model renowned for its efficient and effective feature extraction from raw audio data. The Wav2Vec2 architecture, integral to this model, utilizes self-supervised learning paradigms to capture rich, contextual representations without the need for explicit feature engineering, making it highly suitable for age prediction.

At the core of the model is a classification head termed ModelHead, specifically crafted to process the hidden states generated by Wav2Vec2 and produce an age estimate. This head comprises a linear layer that maps the hidden states to an intermediate dimensionality (equal to the hidden size of the model), followed by a dropout layer to mitigate overfitting. A subsequent linear layer transforms these features into a single predicted age output. Crucially, the activation function employed in the ModelHead is the hyperbolic tangent (tanh), which normalizes the output of the neural network to a range between -1 and 1, facilitating efficient gradient propagation.

During inference, the model processes input audio through the Wav2Vec2 architecture to obtain a series of temporal hidden states. These states are then aggregated, typically by averaging, to condense the information into a single vector that encapsulates the essential characteristics of the input sequence, which is subsequently fed into the age prediction head.

The model hosted on HuggingFace predicts both age and gender, but the gender portion of the prediction has been excluded in this implementation. The authors of the original paper (Burkhardt et al., 2023) boast an MAE of 8.35 years. The model was tested against our dataset (details in above sections) where it achieved an accuracy of 55.6% (+- 6 years tolerance).

Since this approach was the best performing, it was incorporated into the final model for multi-modal age prediction through speech and face image input.

3.3 Fusion Modes

In this section, we delve into various fusion modes employed in multimodal age prediction models. While leveraging separate models for each modality provides flexibility and allows for specialized feature extraction, effectively fusing the outputs of these individual models is essential for capturing complementary information and enhancing prediction accuracy. Fusion modes determine how information from different modalities is combined or integrated to produce the final prediction. By exploring different fusion strategies, ranging from early fusion to late fusion and beyond, we aim to identify the most effective approach for integrating image and audio data in our multimodal age prediction framework.

Point of fusion is a critical consideration in multimodal fusion, determining where in the pipeline the fusion occurs. Late fusion involves processing each modality separately and then combining the results at a later stage. In contrast, early fusion combines raw data from different modalities at the input level, also known as feature-level fusion. Joint fusion integrates modalities throughout the entire pipeline, enabling the model to capture interactions between them. Common space fusion involves mapping data from different modalities into a shared feature space, known as cross-modal embeddings.

For our multimodal age prediction framework, we opt for late fusion due to our previous explorations of models handling the two modalities separately. In late fusion, various fusion modes can be employed to combine the outputs of individual models. Summation fusion involves summing up

or concatenating features from different modalities. Averaging fusion calculates the average output among modalities, smoothing out inconsistencies. Voting fusion employs a feature voting mechanism based on the modality with the highest accuracy. Classifier (also known as weighted fusion) utilizes probability distributions generated by each classifier, with the final decision made by combining these distributions.

4 Final Model

For our final model, we have adopted a weighted average late fusion model with our fine-tuned MobileNetV2 and Wav2Vec2 models.

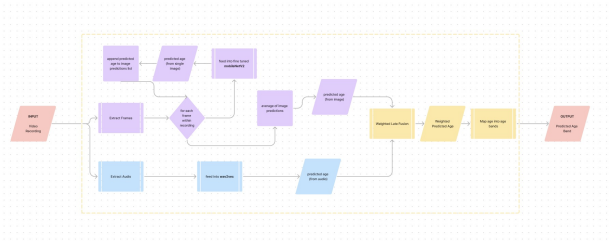

Figure 16: *Final model architecture*

At the lower-level, we have two separate models for the audio and image input – these inputs come from the same video input in which frames and the audio are extracted and fed into its own separate pipelines. With the extracted frames, each frame pre-processed and subsequently fed into the fine-tuned MobileNetV2 model, and the predicted ages are stored in a list which we then average out to get the predicted age from image pipeline. For audio, it is converted into signals and fed into the wav2vec2 model, with its output being the predicted age from audio pipeline. Both of these lower-level models will output an exact age estimation.
At a higher level these two outputs from the separate modality pipelines are fused with a weighted average late fusion, in which its weight is accounted for based on the separate models' relative accuracy. This produces a weighted predicted age which is then mapped into age bands – giving us the final output.

4.1 User Interface

A simple UI was implemented which displays the predicted age-band on top of the user's detected face. On the screen are also instructions on how to use the program and the separate results from the audio model and image model. Two random sentences are given to the user to say. Libraries used are pyaudio, cv2, and wonderwords.

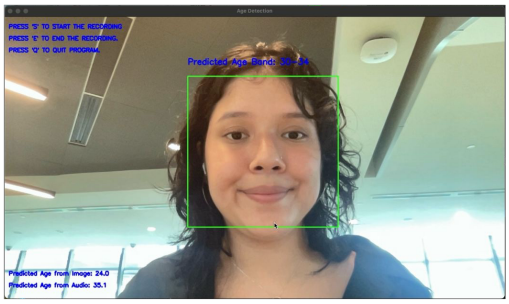

Figure 17: *UI implementation*

5 Results

As our final model was a fusion of two separate models, it is evaluated by weighing out the results of the two models. Fused model accuracy is calculated by weighing out the relative normalized accuracy:

$$0.566 \div (0.566+0.637) + 0.637 \div (0.566+0.637) = 0.599554$$

Model	Accuracy
wav2vec 2.0	55.6%
MobileNetv2	63.7%
Fused Model	59.96%

Table 10: *Results of final model*

However, this accuracy is deemed theoretical as our final fused model has not been sampled against a dataset for direct accuracy measurements. In this calculation, we forgo the fact that the fused model accounts for possible volatility in age predictions, from the image model, by averaging out the multiple frames taken from video input. This might lead to a higher predicted accuracy when measured against our fused model implementation.

6 Discussion

The attainment of a 59.96% theoretical accuracy in multimodal age prediction models using video input carries implications for various use-cases in which age prediction can be utilized.

While achieving a theoretical accuracy of 59.96% marks an accomplishment in the development of multimodal age prediction models from video input, it's essential to acknowledge the limitations posed by the modest accuracy level. The model's performance might still fall short of real-world expectations, particularly in scenarios where precise age estimation is critical, such as age-restricted content access or age-sensitive healthcare interventions. Moreover, inaccuracies in age prediction could potentially lead to misclassifications and subsequent implications in decision-making processes, like personalized content recommendations or targeted advertising campaigns. It is also important to note that the reliability of age verification mechanisms in ensuring online safety may be compromised if the model fails to accurately distinguish between age groups. Therefore, while the model's performance represents progress, addressing its limitations and striving for higher accuracy levels remain imperative for realizing its full potential across various applications.

7 Future Improvements

7.1 Refining Transfer Learning on Models

Transfer learning involves leveraging knowledge from pre-trained models on large datasets to solve related tasks with smaller datasets. Therefore, there can be various strategies employed to further refine transfer learning for age prediction specifically. One of which is fine-tuning layers of pre-trained models while keeping others frozen. This allows the model to retain important features learnt during pre-training while adapting to specific characteristics of the prediction task.

Another approach would be to do systematic hyperparameter tuning, including parameters such as learning rate, batch size and regularization strength which is essential for optimizing the transfer learning approach. Experimenting with various model architectures or variations like incorporating regularization techniques to prevent overfitting are vital for refining an age prediction model.

7.2 Class Imbalance in Audio Data

Curating a balanced dataset of speech, which is varied in accents, gender, and of sizeable duration would help in the training of audio models immensely.

8 References

Datasets

CheyneyComputerScience. (n.d.). GitHub - CheyneyComputerScience/CREMA-D: Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D). GitHub. https://github.com/CheyneyComputerScience/CREMA-D/tree/master

JingchunCheng. (n.d.). GitHub - JingchunCheng/All-Age-Faces-Dataset: All-Age-Faces (AAF) database. GitHub. https://github.com/JingchunCheng/All-Age-Faces-Dataset

UTKFace. (n.d.). UTKFace. https://susanqq.github.io/UTKFace/

James et al., (2016, Zenodo - Children speech recording (English, spontaneous speech + pre-defined sentences). Zenodo. https://zenodo.org/records/200495

Kallay, J., & Redford, M. A. (2020). Clause-initial AND usage in a cross-sectional and longitudinal corpus of school-age children's narratives. Journal of Child Language, 48(1), 88–109. https://doi.org/10.1017/s0305000920000197

Speech Accent Archive. (2017, November 6). Kaggle. https://www.kaggle.com/datasets/rtatman/

speech-accent-archive

Population and population structure. (n.d.). Base. https://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/latest-data

Team, K. (n.d.). Keras documentation: MobileNet, MobileNetV2, and MobileNetV3. https://keras.io/api/applications/mobilenet/

DenseNet. (n.d.-b). https://huggingface.co/docs/timm/en/models/densenet#how-do-i-train-this-model

Models

audeering/wav2vec2-large-robust-24-ft-age-gender · Hugging Face. (n.d.). https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender

Openai/clip-vit-base-patch32 · hugging face. openai/clip-vit-base-patch32 · Hugging Face. (n.d.). https://huggingface.co/openai/clip-vit-base-patch32

Kundu, N. (2023, January 23). Exploring ReSNet50: An In-Depth look at the model architecture and code implementation. Medium. https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at\-the-model-architecture-and-code-\implementation-d8d8fa67e46f

Other Literature

Burkhardt, F., Wagner, J., Wierstorf, H., Eyben, F., audEERING GmbH, Chair EIHW, University of Augsburg, & GLAM, Imperial College London. (2023). Speech-based Age and Gender Prediction with Transformers. Abstract. https://arxiv.org/pdf/2306.16962.pdf