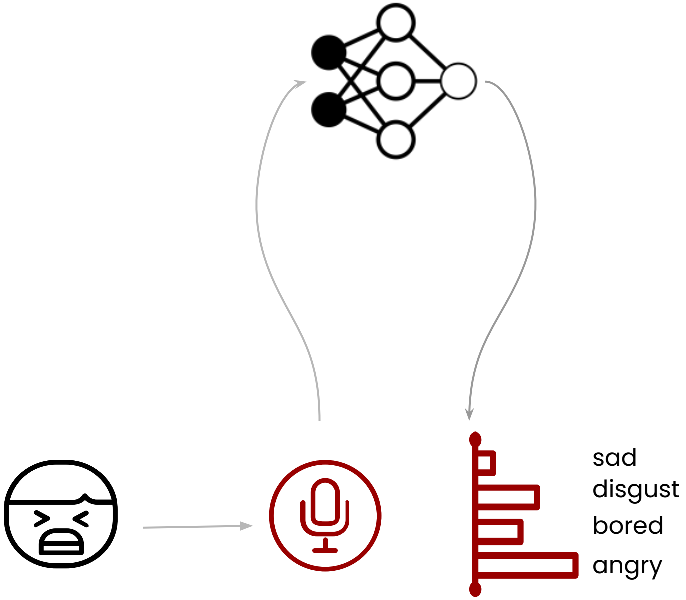
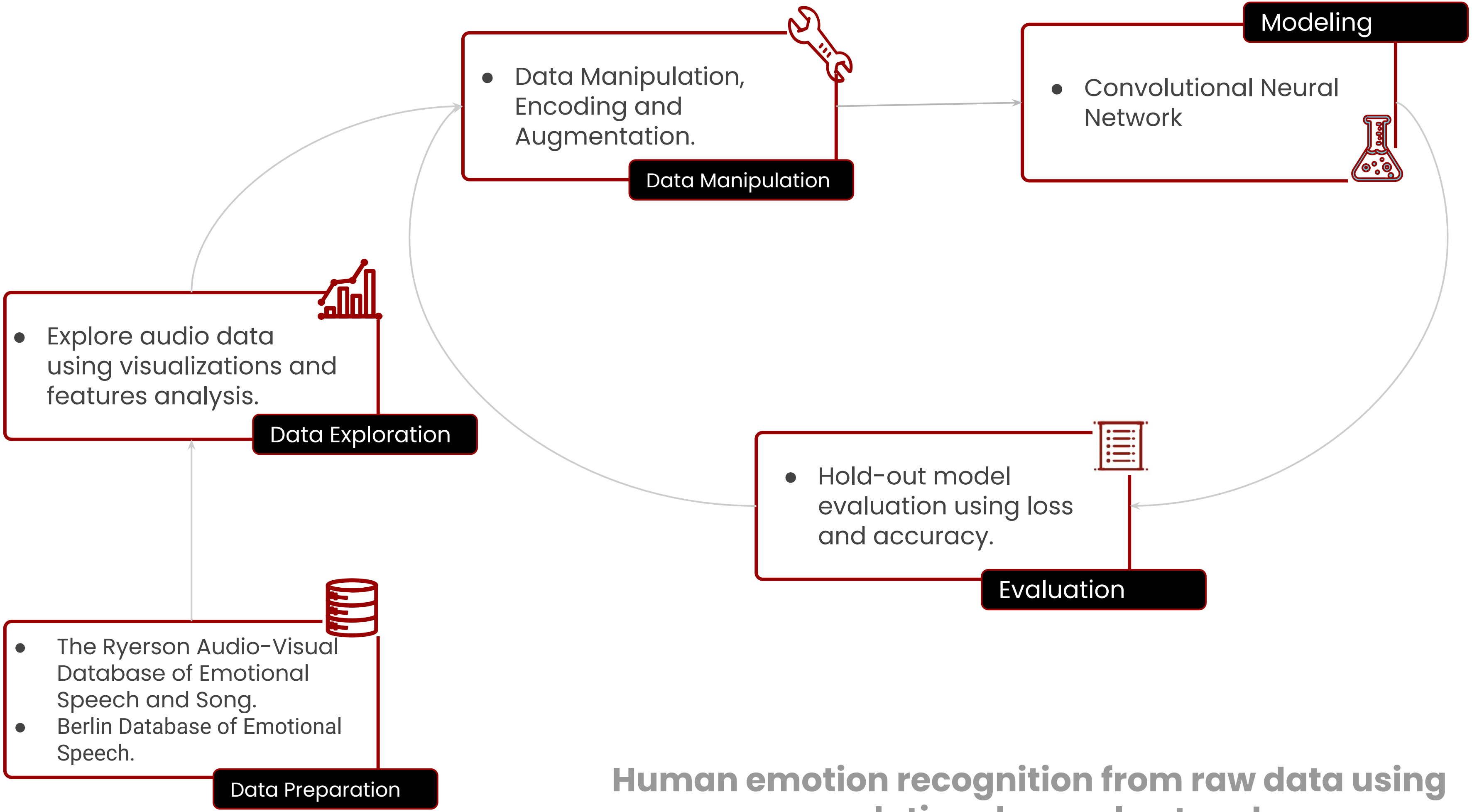
Emotions Recognition from Audio Speech Using Deep Learning

Pattern Recognition - F20

Fatima AlSaadeh



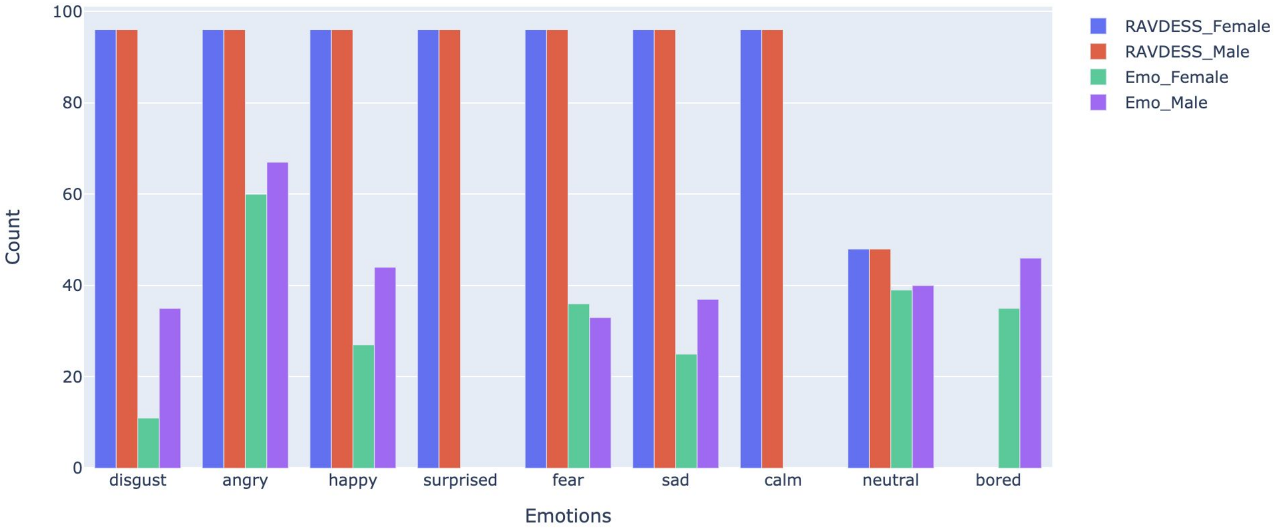


Human emotion recognition from raw data using convolutional neural networks

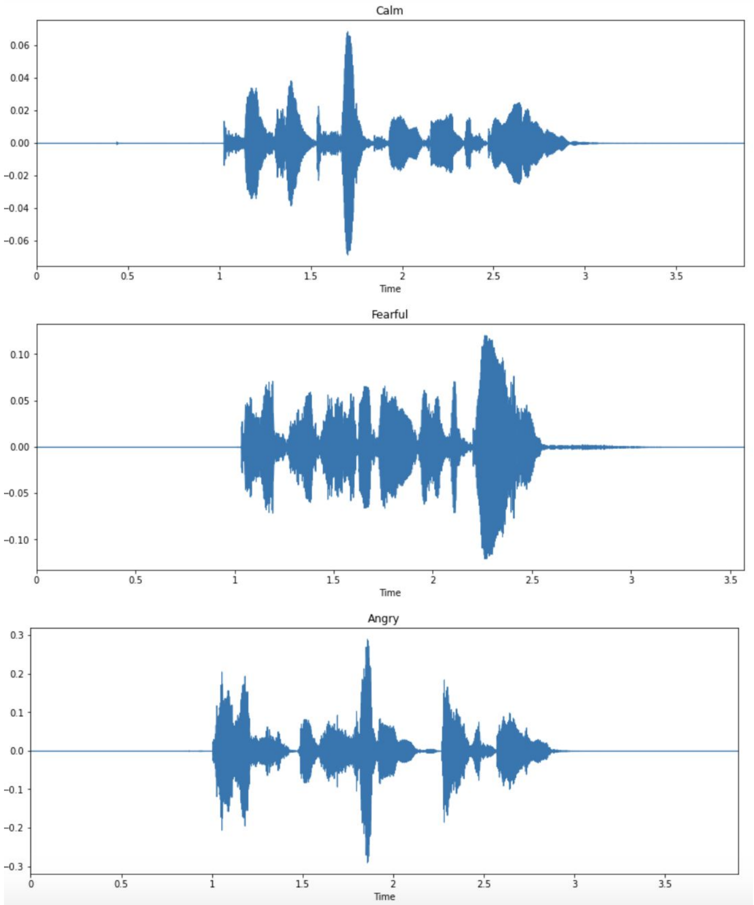
Data Preparation

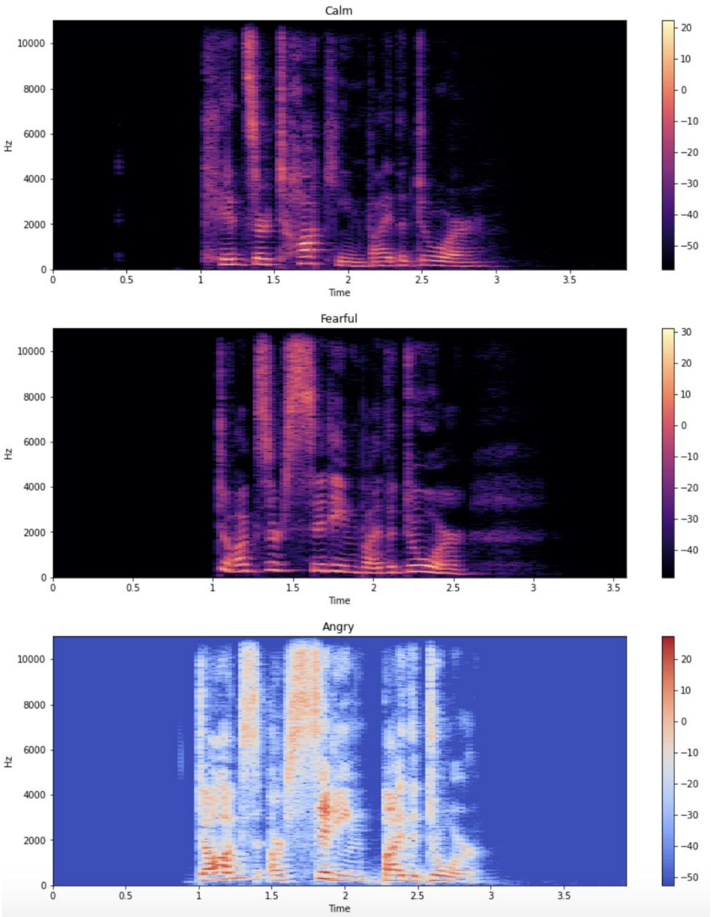
1. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS):
- The audio-only files contain 1440 files, 24 actors, 12 male and 12 female.
- English.

2. Berlin Database of Emotional Speech (EMO-db) :
- The audio files contain 535 files, 302 males, 233 females
- Germany



Data Exploration

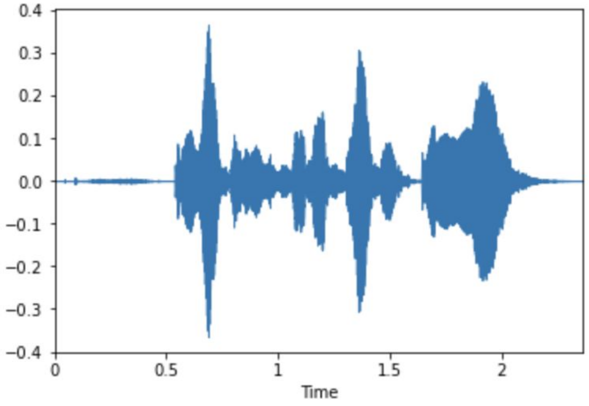




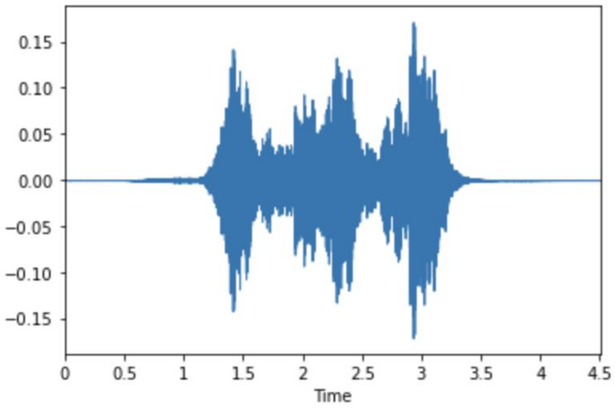
Data Manipulation

- Data Standardization

$$x' = \frac{x - \bar{x}}{\sigma}$$



- Data Augmentation
  - Adding white noise.
  - Stretching the sound.
  - Random Shifting.



- Resampling and Reduction.

- Classes one-hot encoding.

Convolutional Neural Networks

- Require minimal data pre-processing, due to their convolutional layers and their ability to extract features, eliminating the feature engineering by hand step.

$$x_i^l = \sum_{n=1}^{j} w_{i,j} * x_i^{l-1} + b_i^l$$

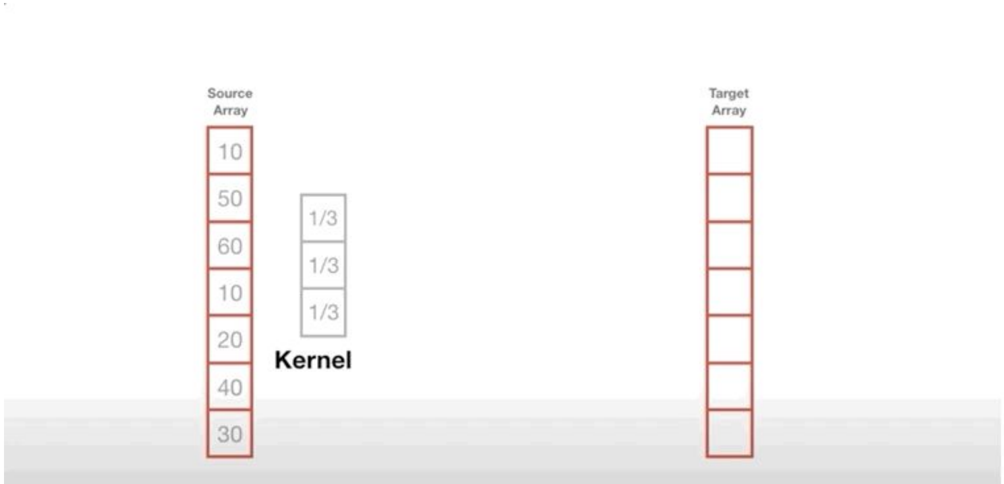
- It takes the raw data as an input and built of different convolutional, pooling and fully connected layers.
- Convolutional Layers have filters which help detect the patterns in the raw input data.

Model Architecture

- Input
- Temporal Convolution: (Conv1D)

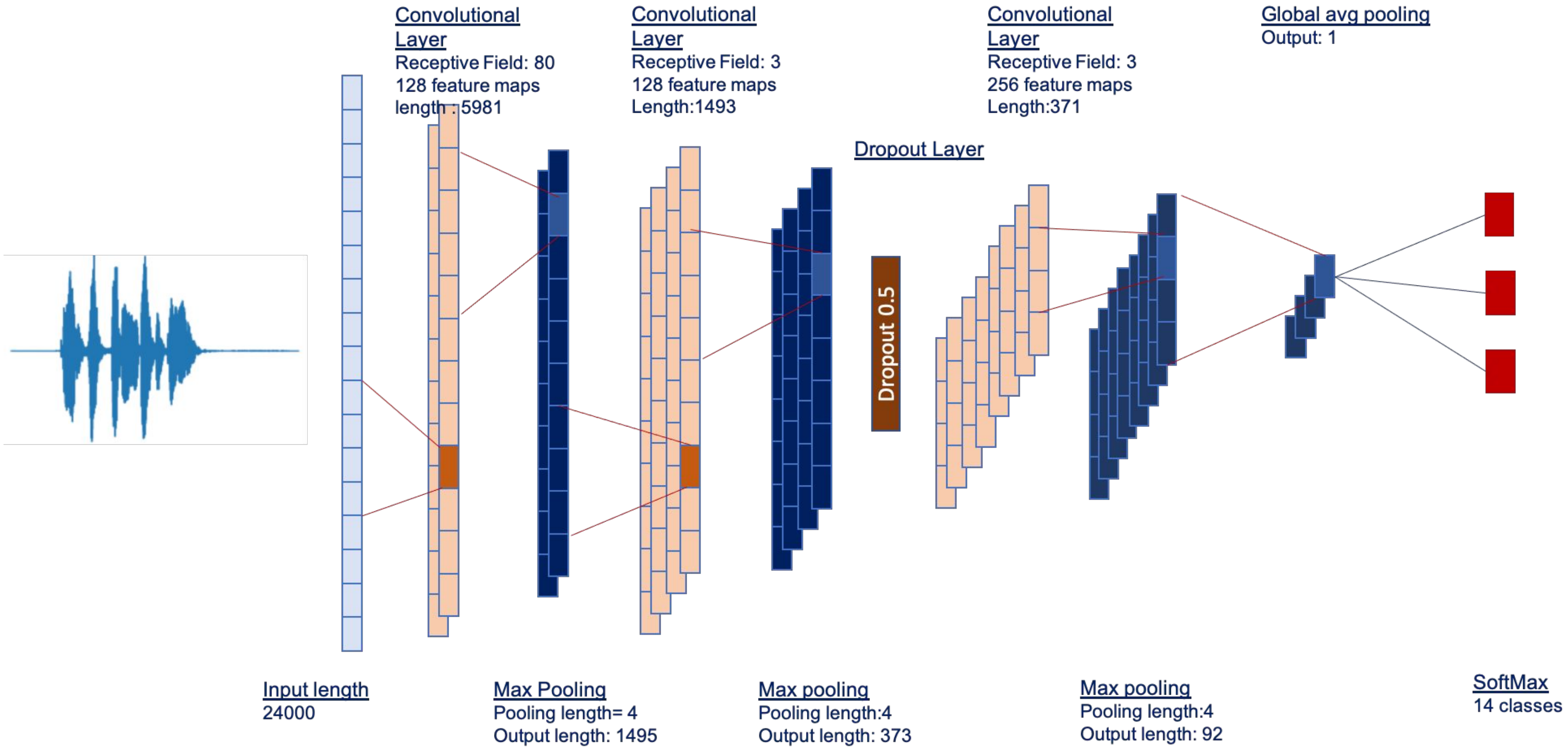
$$Y = (X - F + 2*P)/S + 1$$

- Batch Normalization.
- Max Pooling .
- Activation function.
- Dropout.
- Average Pooling
- Softmax.



Credit: https://www.youtube.com/watch?v=ulKbLD6BRJA

Model Architecture



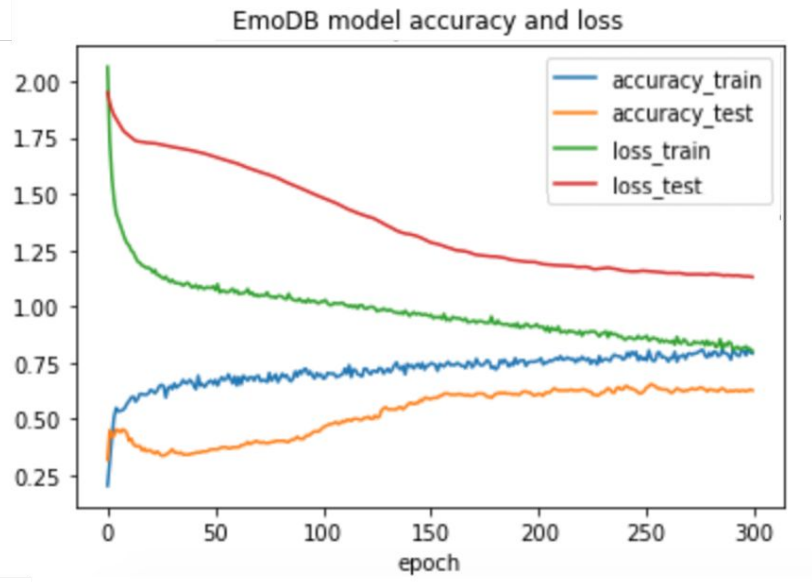
Evaluation

- **Evaluation metrics:**
  - Accuracy.
  - Precision.
  - Recall.
  - Loss

- **Fitting and testing the model to predict the classes in different categories:**
  - 2 classes : Emotions Intensity strong and natural.
  - 4 classes: positive, negative, fearful and surprised.
  - 7 classes emotions after we merged the neutral and calm.
  - 14 classes: all emotions male and female: male and female, neutral - calm, happy, sad, angry, fearful, disgust, surprised.

- **Using holdout evaluation method:**
  - Split the data into training and testing datasets 80%, 20%
  - Further split the training data into training and validation 80%, 20%.

Results Analysis



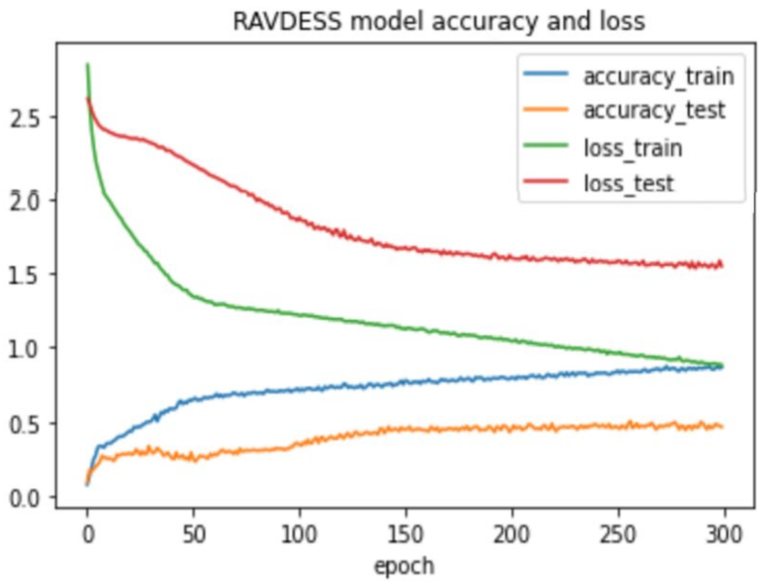


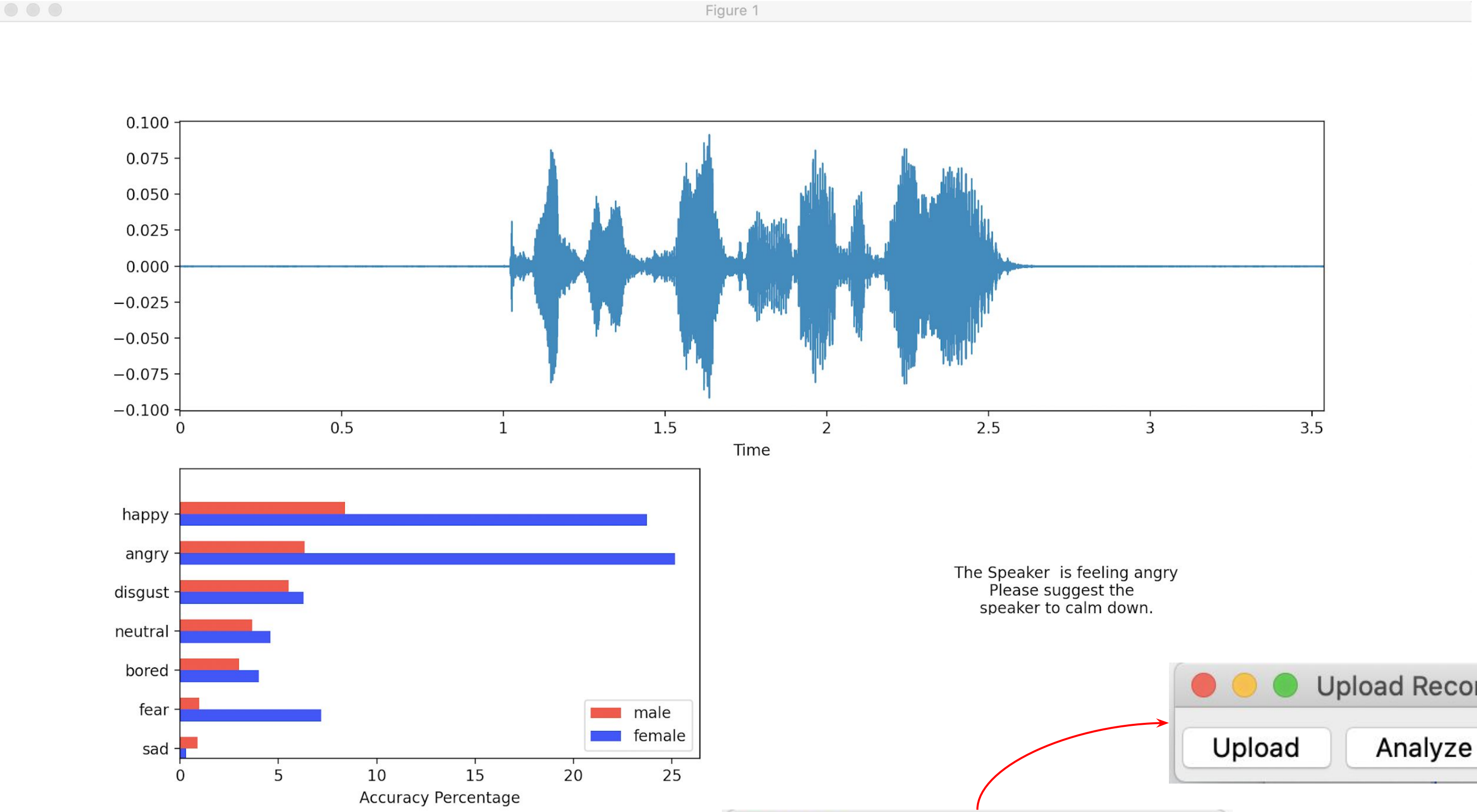
Table 2. Train and test accuracy - EmoDB

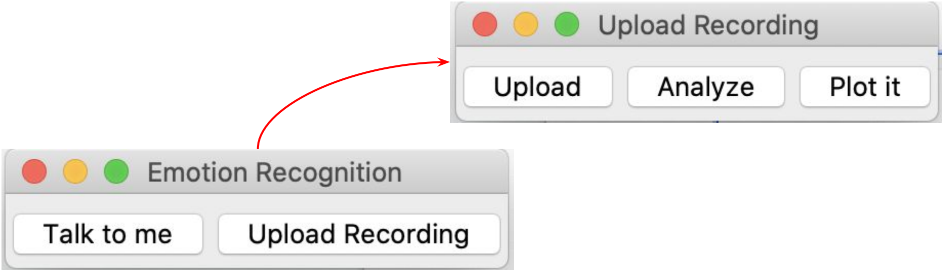
Classes	Train	Test
4-classes	87.1%	71.03%
7-classes	85.9%	60.2%
14-classes	87.9%	60.7%

Table 1. Train and test accuracy - RAVDESS

Classes	Train	Test
2-classes	89.0%	68.0%
4-classes	91.8%	55.7%
14-classes	86.5%	51.8%

Applications





References

[1] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A.Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. 2016

[2] K. Venkataramanan and H. R. Rajamohan. Emotion recognition from speech.CoRR, abs/1912.10458, 2019

[3] J. Rintala. Speech Emotion Recognition from Raw Audio using Deep Learning. 2020

[4] W. Dai, C. Dai, S. Qu, J. Li, and S. Das.Very deep convolutional neural networks for raw waveforms.CoRR,abs/1610.00087, 2016

[5] Livingstone sr, russo fa (2018) the ryerson audio-visual database of emotional speech and song (ravdess): A dynamic,multimodal set of facial and vocal expressions in north american english. plos one 13(5): e0196391

[6] Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, andB. Weiss. A database of german emotional speech. volume 5,pages 1517–1520, 01 2005.