Team Genesis

Classifying Emotions from Audio

The Erdős Institute, May 2022 Bootcamp

Mario Gomez Flores, Tajudeen Mamadou, Mohammad Nooranidoost, Elif Poyraz, Rose Weisshaar

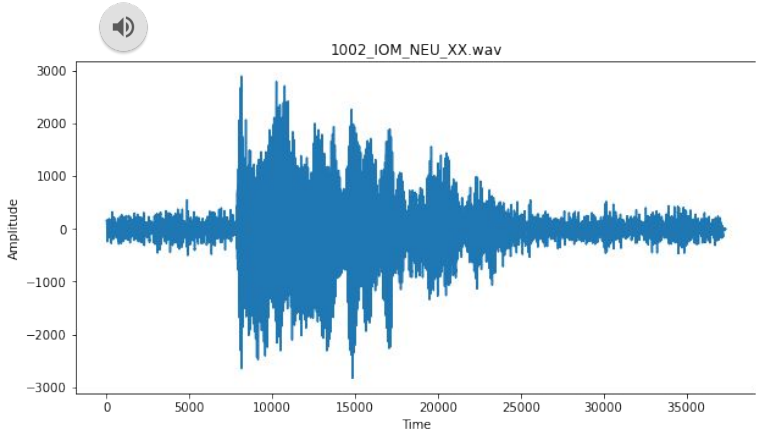
https://github.com/elfnrpyrz/erdos-may22-genesis

I. Problem and Stakeholders

Problem:

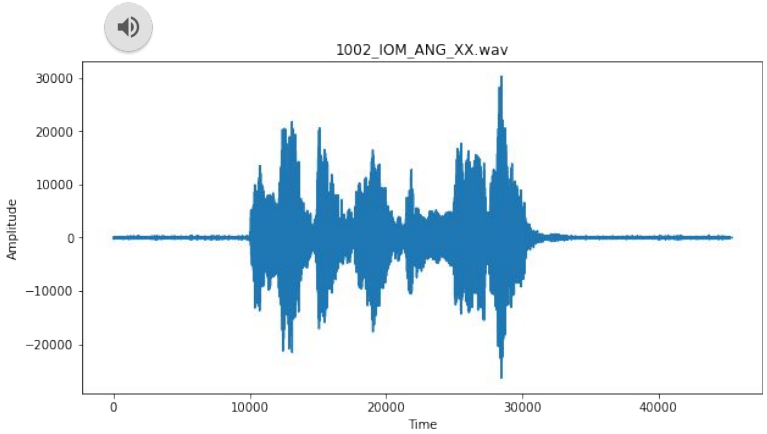
- Classify the emotional content of human speech.

Example:



Stakeholders:

- Children on the autism spectrum
- Language translation
- Speech-to-text



II. Dataset & Measures of Success

The CREMA Dataset

- 6076 audio files
  - 92 Actors
  - 11 sentences
  - 6 emotional categories:
    - Neutral, Anger, Happy, Sad, Fear, Disgust.
- Crowd-sourced data for human listeners
  - Participants tried to identify the intended emotion of an audio clip
  - Each clip has at least 10 raters.

Measures of Success

- CREMA:
  - Human listeners accurately classified 40% of audio files.
- Our measure of success:
  - Model accurately classifies at least 40% of audio files.

CREMA-D: https://github.com/CheyneyComputerScience/CREMA-D

III. Workflow

Data Clean Up ***Feature Extraction*** ***Model Training*** ***Model Performance***

Remove Noise
Remove Silence

Audio
Mid-level Features

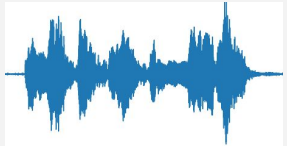
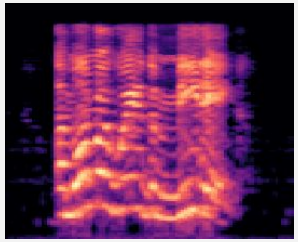


Image
Spectrograms

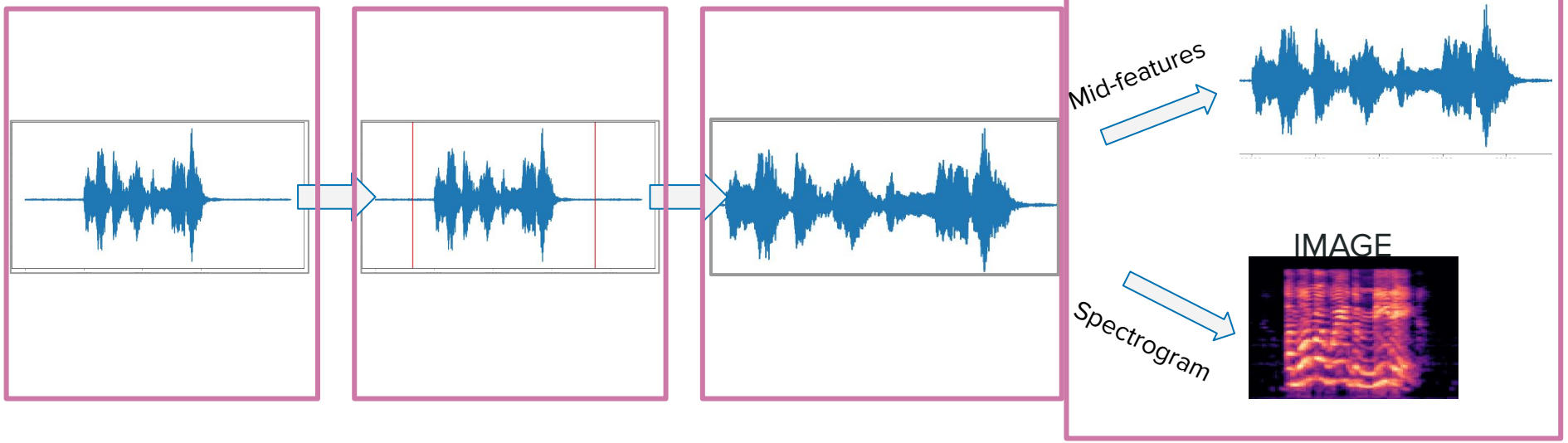


SVM

CNN

Accuracy of classification
Comparison to Human Raters

IV. Data Clean Up

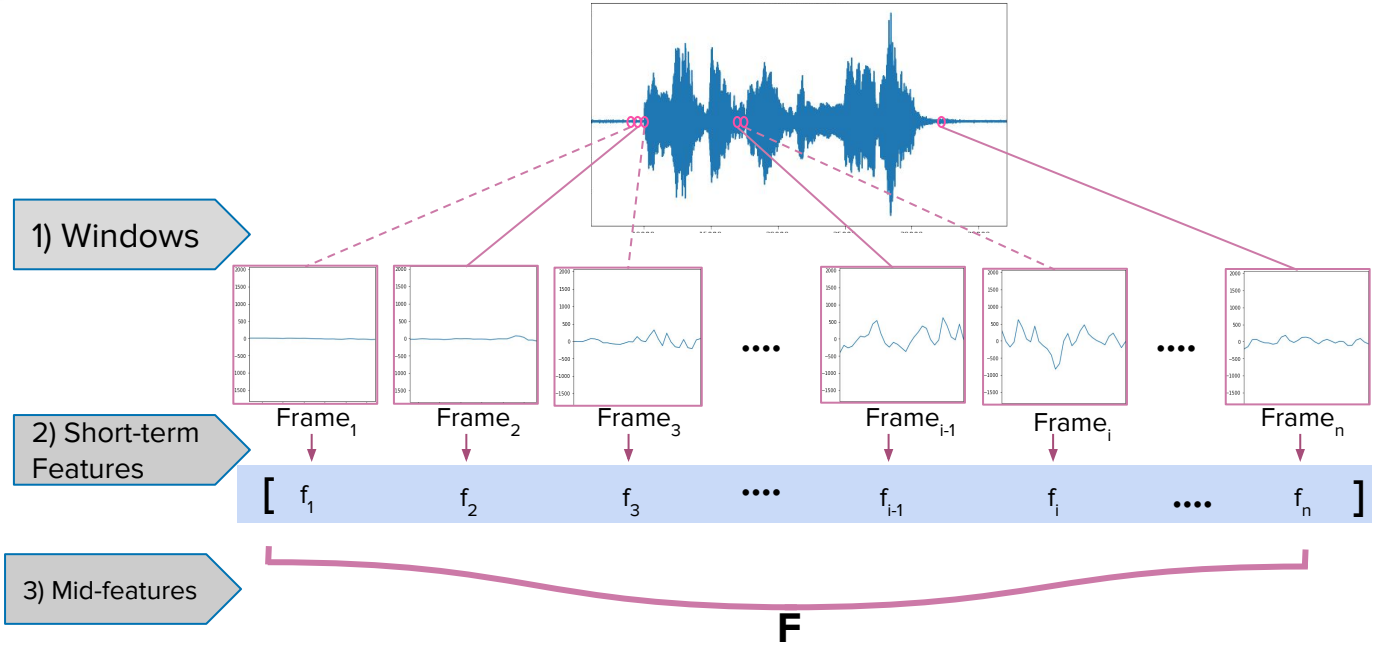


V. Feature Extraction

Mid features

- Python Package: **PyAudioAnalysis**

- Advantages:
  - Interpretability
  - Low computational cost
- Disadvantages:
  - Lost time information

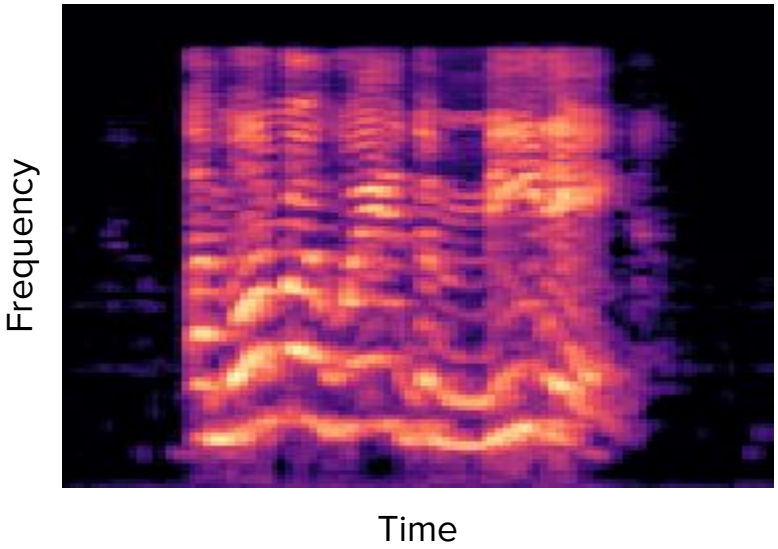


V. Feature Extraction

Spectrograms

- What is a spectrogram?
  - Color represents intensity
- Python Package:
  Librosa

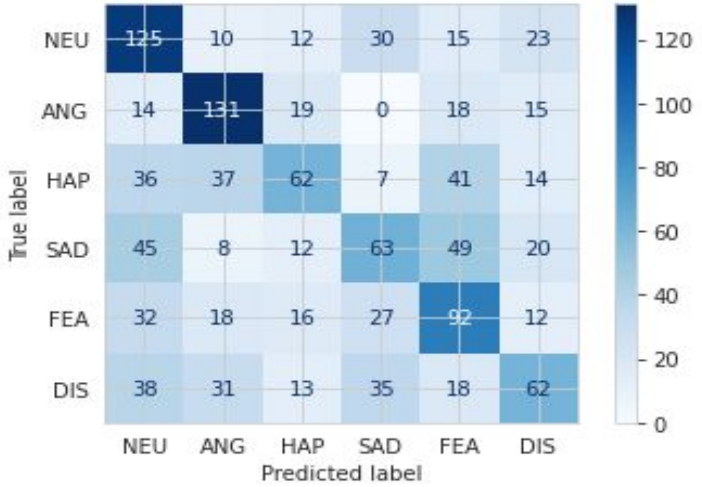
- Advantages:
  - Uniform size
  - Apply image-processing techniques
- Disadvantages:
  - Interpretability
  - Computation time



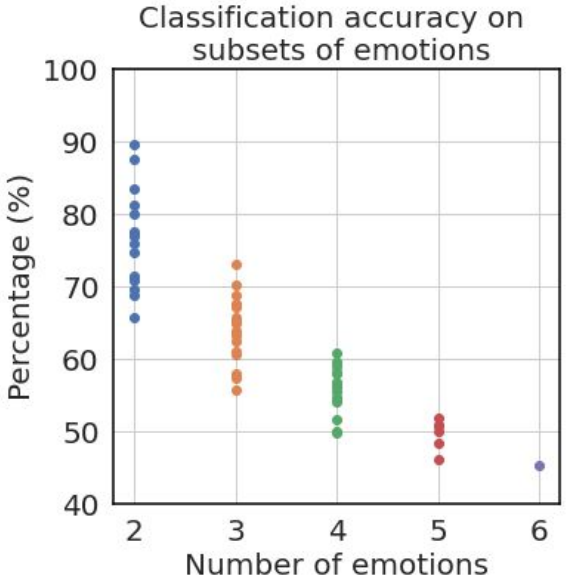
VI. Models

Model 1: Support Vector Machine

- Runs on mid features
- PCA did not help
- Overall accuracy: 45%
- Human rater accuracy: 40%



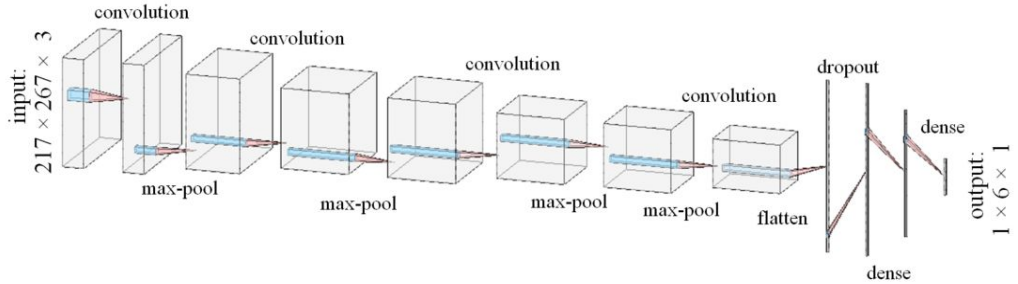
- Experimented with different subsets of emotions

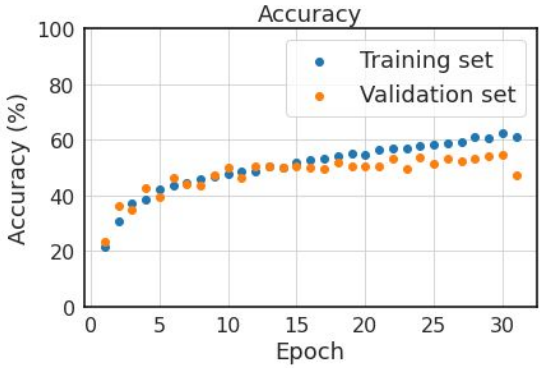


VII. Models

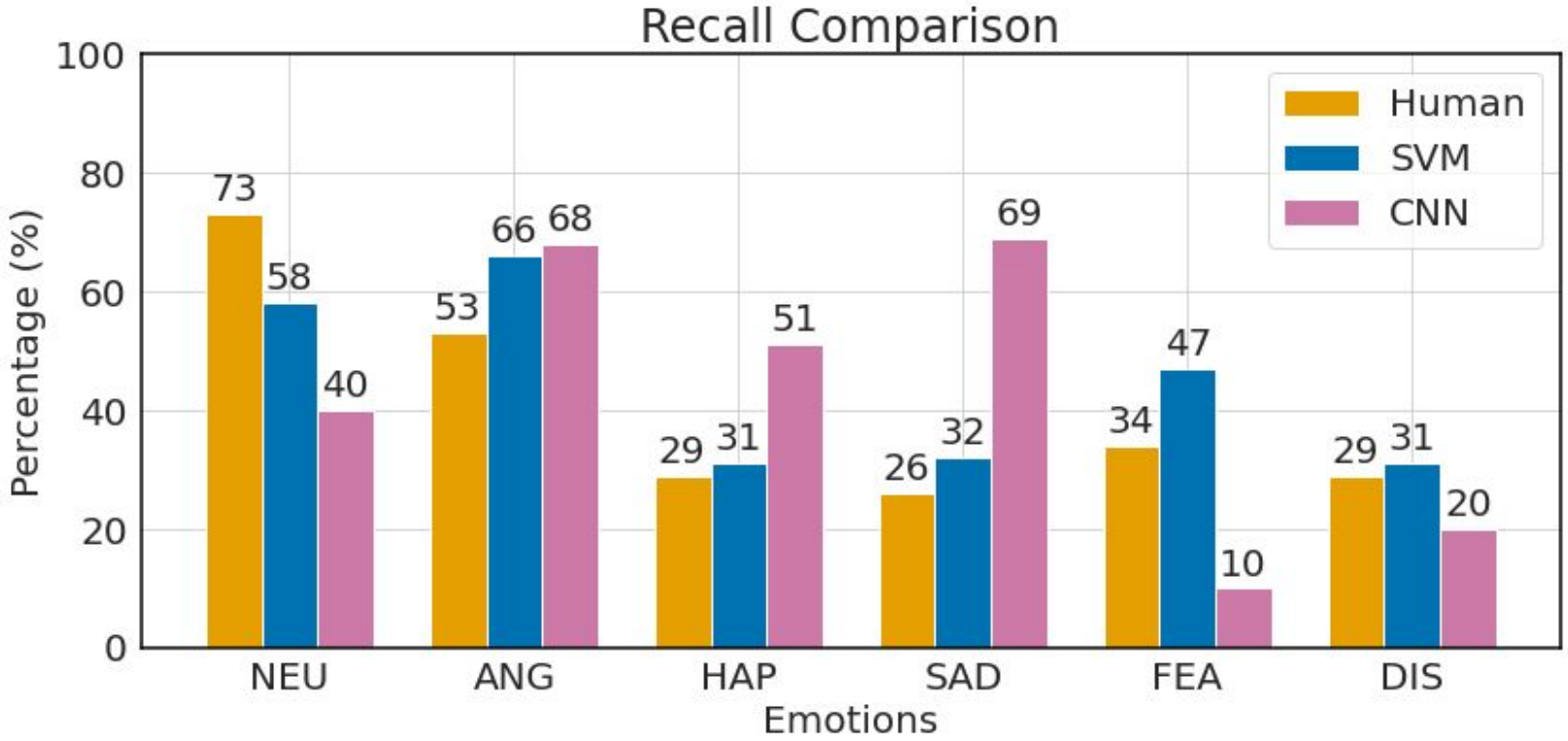
Model 2: Convolutional Neural Network

- Runs on spectrograms
- To combat overfitting:
  - Data augmentation
  - Dropout layer
- Training set size: ~4900
- Validation accuracy: ~50%
- Test set accuracy: 42%
- Human rater accuracy: 40%





VIII. Model Evaluation



IX. Summary & Future Directions

Summary

- Two models with human-like accuracy
- SVM model was more similar to human performance

Future Work

- Train on a new dataset with labels from human listeners
- Extract mid features for each word
- Naturally occurring data (e.g. phone conversations)

We are grateful to the authors who developed **the CREMA dataset**.

We are also grateful to our Mentor **Akul Dewan**, and **the Erdős Institute Team**!

Personal Links

https://www.linkedin.com/in/elfnrpyrz/

https://www.linkedin.com/in/mario-gomez-flores

http://nooranidoost.com/

www.linkedin.com/in/tajudeen-mamadou

https://math.wfu.edu/people-faculty