SER: Speech Emotion Recognition using Learning Algorithms

Gabriel Cha , Arman Rahman, Jun-Hee Hwang
gcha@ arahman@ juh016@

Mentor: Justin Eldridge
jeldridge@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

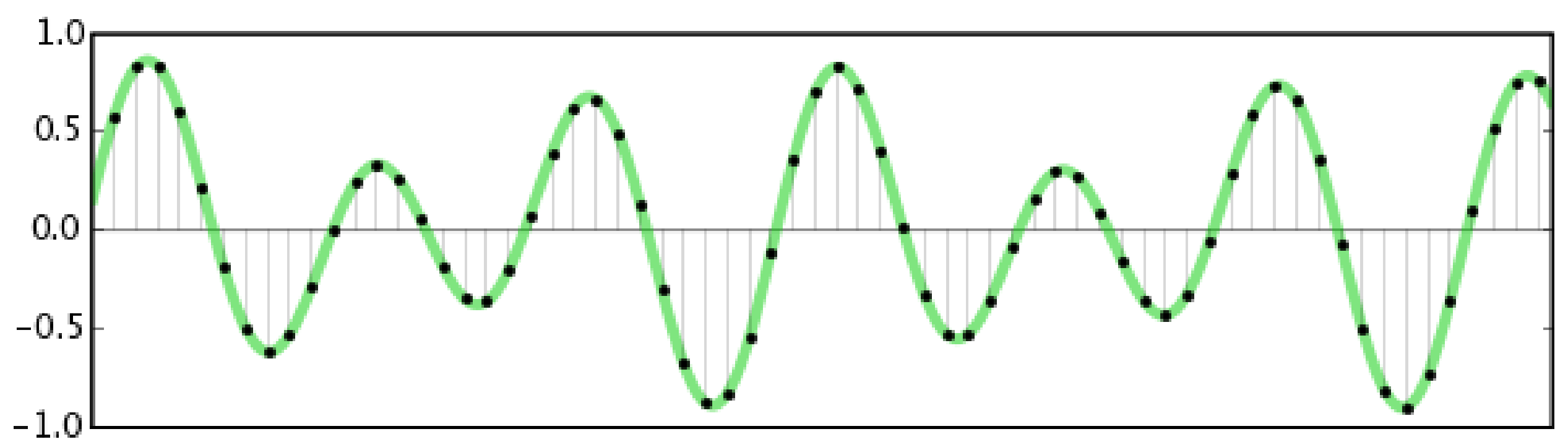
Motivation

Emotion is a crucial aspect of human-computer interaction (**HCI**), influencing how users perceive and interact with technology.

By enabling intelligent systems to recognize and respond to emotions, we can enhance user experience, making interactions more engaging.

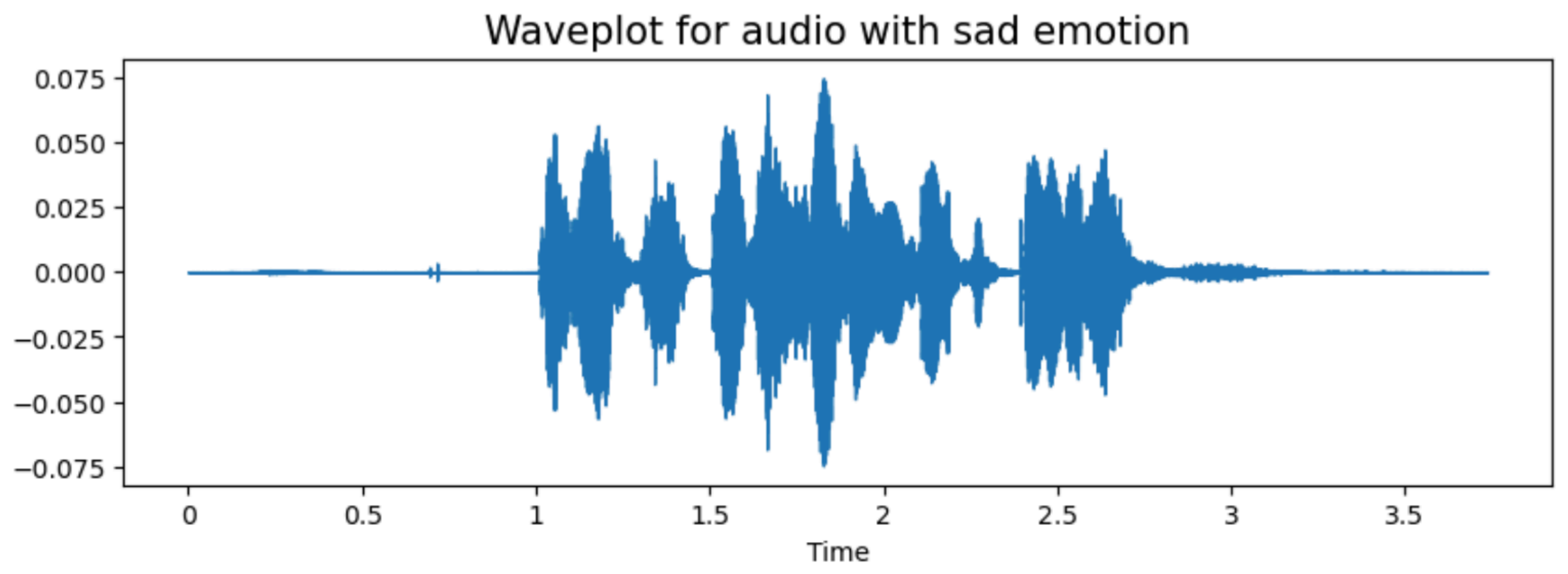
Objective

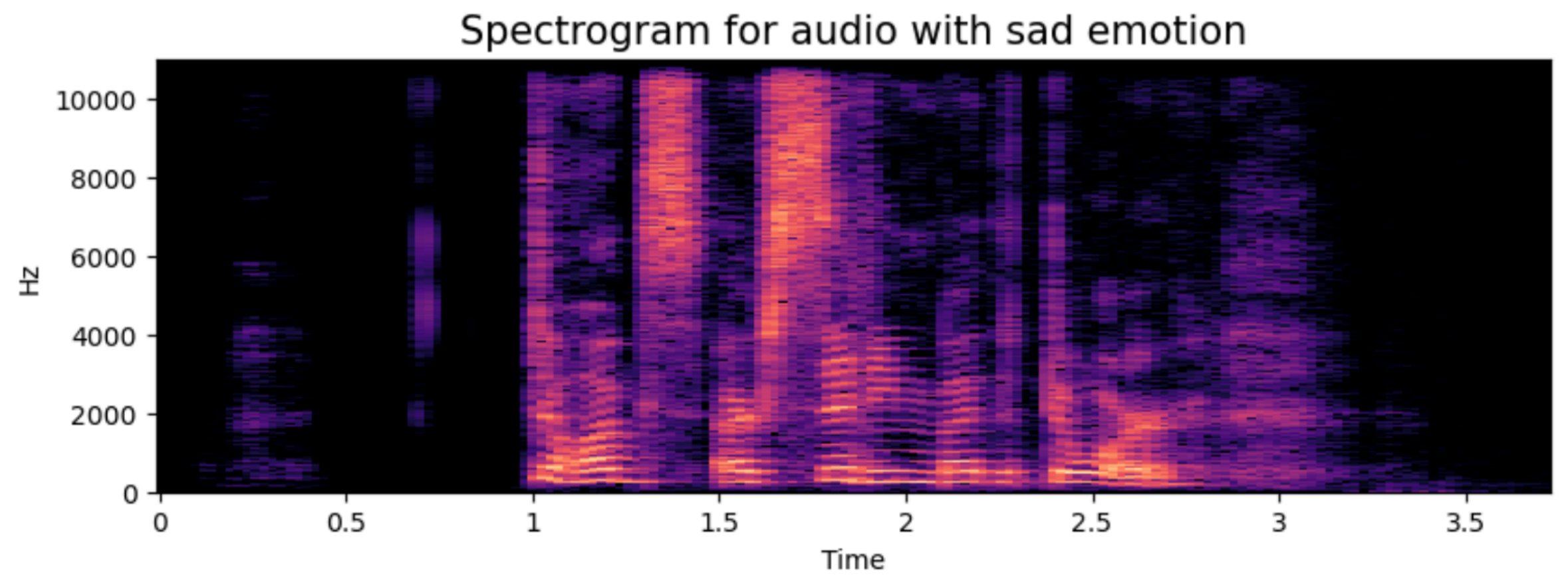
Our study aims to classify emotions using machine learning algorithms and compare the results of Random Forest, SVM, CNN, and ViT.



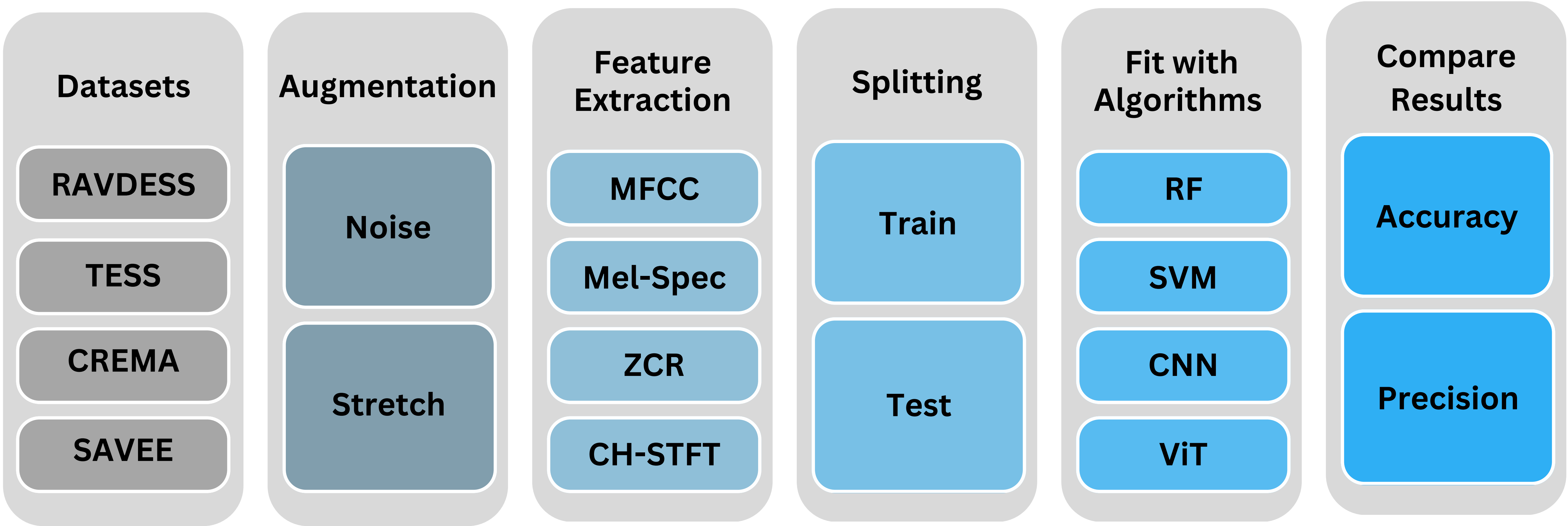
Challenges: Human emotions are expressed with varied and complex differences in tone, pitch, volume, speed, and articulation. Effective feature extraction techniques are essential to accurately capture these nuances.

Methodology


Waveplot for audio with sad emotion


Spectrogram for audio with sad emotion

Mel-Spectrogram: Decomposes an audio signal into short time frames and applies the Short-Time Fourier Transform (STFT) to obtain the amplitude of various frequencies in each frame.



Datasets: RAVDESS, TESS, CREMA, SAVEE

Augmentation: Noise, Stretch

Feature Extraction: MFCC, Mel-Spec, ZCR, CH-STFT

Splitting: Train, Test

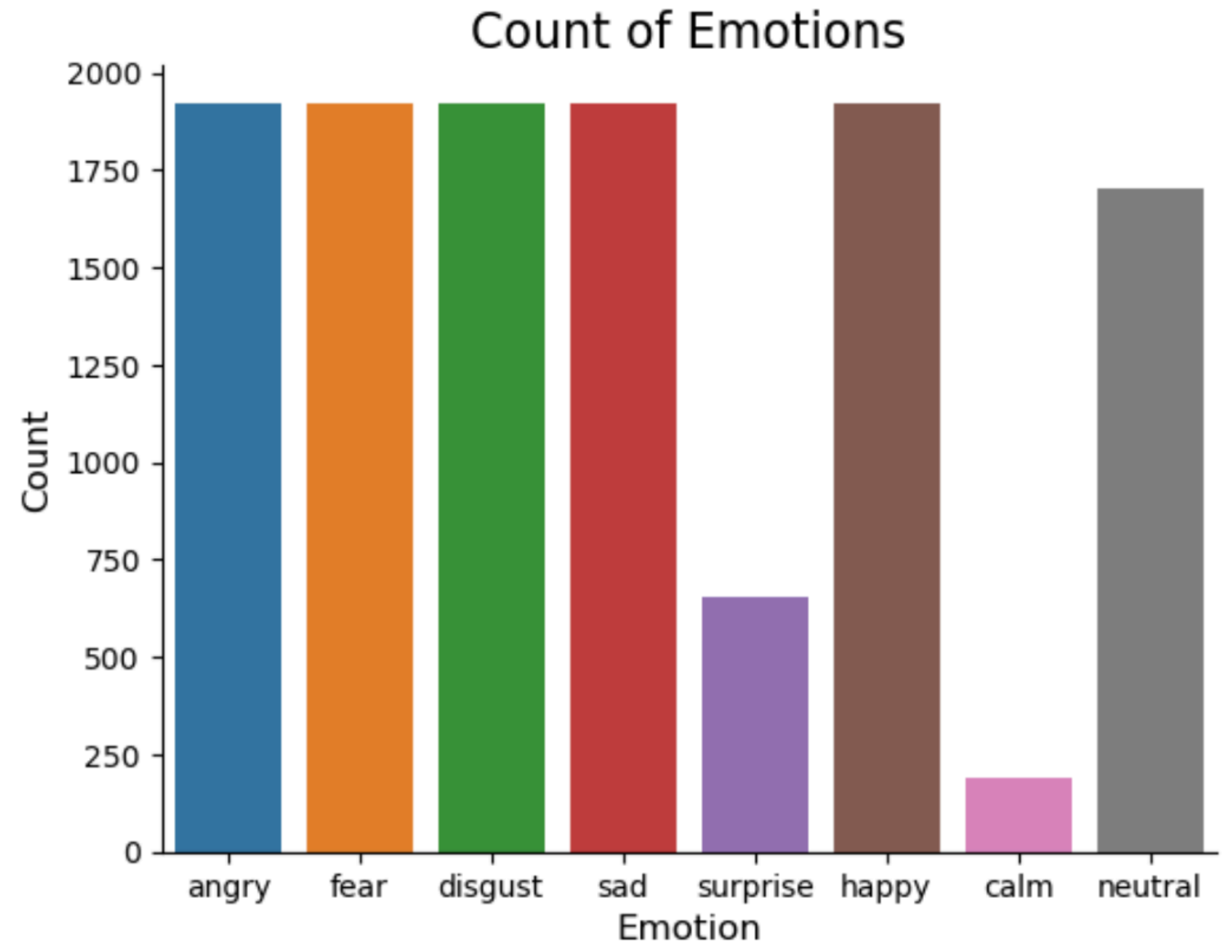
Fit with Algorithms: RF, SVM, CNN, ViT

Compare Results: Accuracy, Precision

Data Collection: Aggregated 12,162 audio clips of voice actors in controlled laboratory settings from the following repositories: RAVDESS, TESS, CREMA, and SAVEE .

Feature Extraction:

- **Zero Crossing Rate** (ZCR): This is the frequency at which the signal crosses zero.
- **Mel-Frequency Cepstral Coefficients** (MFCC): Mapping Fast Fourier Transform (FFT) onto mel scale and applying logarithmic compression.
- **Chroma-STFT** (CH-STFT): Fourier analysis across entire signal, representing pitch distribution.


Count of Emotions

Evaluation

Model	Precision	Accuracy
CNN	78.310183	75.992107
SVM	78.273573	73.722868
Random Forest	74.437301	71.793466
Vision Transformation	73.868650	68.762844
E CNN	57.934682	57.371882

Conclusion & Discussion

Performance Gap: There is a noticeable gap between top-performing models (CNN and SVM) and RF, suggesting the task benefits from models that capture spatial hierarchies and patterns.

Data Insufficiency: The E-CNN lower accuracy could stem from using separate models for male and female speakers, resulting in insufficient data for each CNN to learn intricate patterns.

Feature Extraction: Experiments revealed some features are more influential in classification.

References

We thank our mentor Justin Eldridge for his guidance and support on this project.

[1] Khalil R.A., Jones E., Babar M.I., Jan T., Zafar M.H., Alhussain T. Speech emotion recognition using deep learning techniques: A review.

[2] Singh J, Saheer LB, Faust O. Speech Emotion Recognition Using Attention Model. Int J Environ Res Public Health.