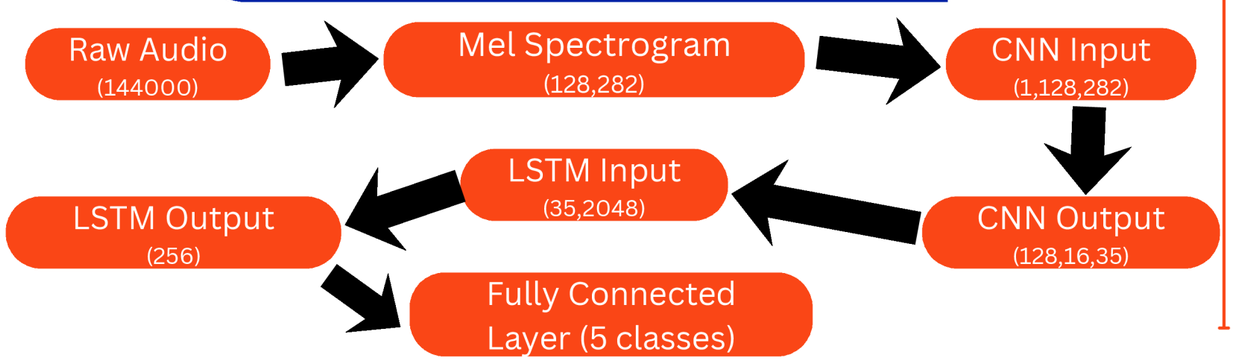
SPEECH EMOTION CLASSIFICATION SYSTEM USING RANDOM FOREST AND CNN-LSTM MODELS

Authors Haichen Fan, Kuan-Chen Chen, Zixuan Liu **Affiliations** Fundamentals of Machine Learning

INTRODUCTION

- Human speech conveys rich emotional information beyond words.
- This project develops a machine learning system that classifies five emotions — happy, sad, angry, surprised and neutral — from short audio clips.
- We compare classical machine-learning (Random Forest) and deep-learning (CNN-LSTM) approaches using a custom audio clip dataset.

CNN-LSTM IMPLEMENTATION



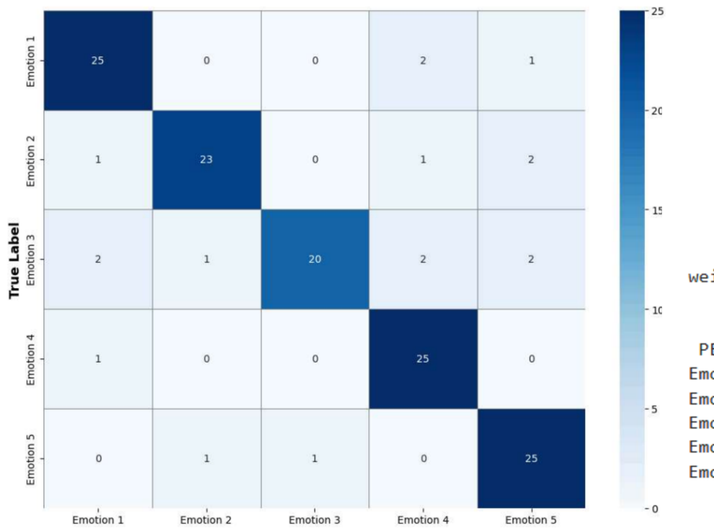
OBJECTIVE

Design and evaluate a robust speech-emotion classifier and compare the effectiveness of baseline and deep-learning architectures on small custom datasets.

METHODOLOGY

- Recorded 3-second WAV clips among five emotion categories.
- Applied augmentation (pitch shift, noise addition, time stretch) → expanded to 36,096 samples.
- Extracted MFCCs, pitch, energy, zero-crossing rate, and tempo features.
- Random Forest trained on engineered features.
- CNN-LSTM trained on Mel-spectrograms using cross-entropy loss.
- Train/validation/test split enforced strict speaker separation.

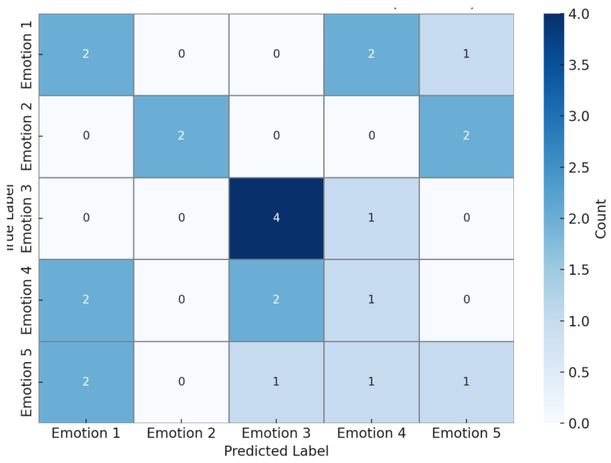
RESULTS/CONCLUSION



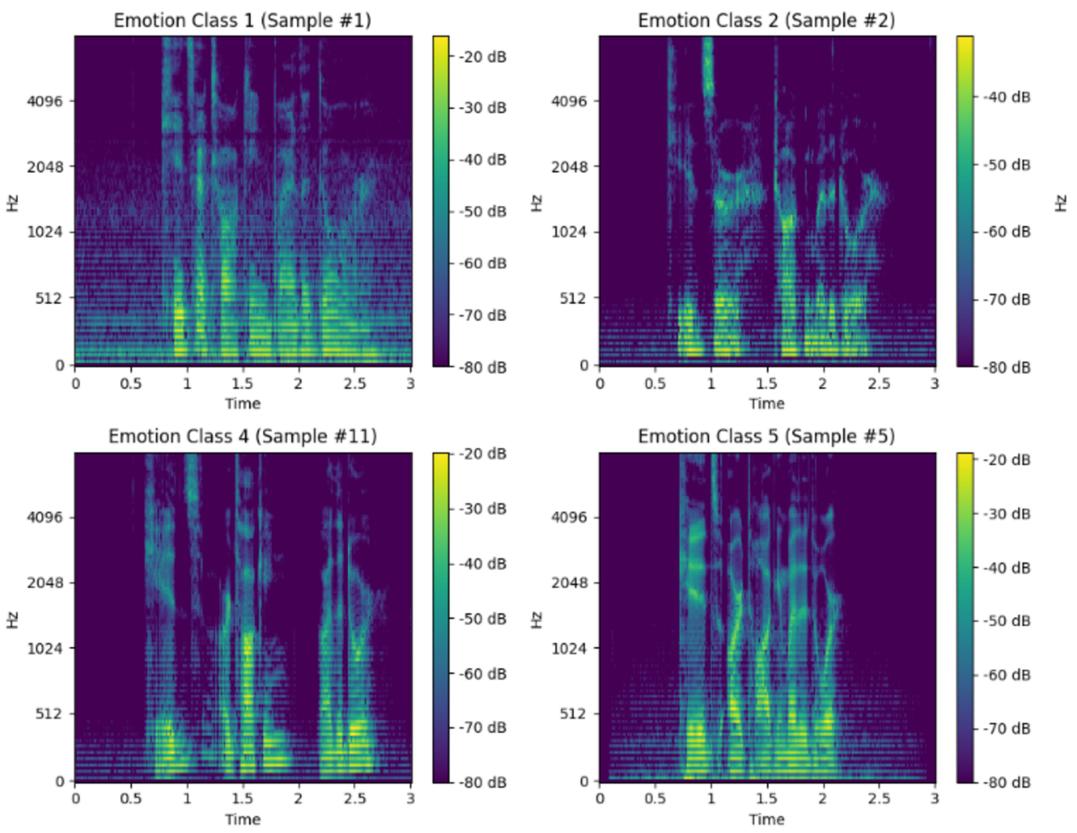
	precision	recall	f1-score	support
Emotion 1	0.8621	0.8929	0.8772	28
Emotion 2	0.9200	0.8519	0.8846	27
Emotion 3	0.9524	0.7407	0.8333	27
Emotion 4	0.8333	0.9615	0.8929	26
Emotion 5	0.8333	0.9259	0.8772	27
accuracy			0.8741	135
macro avg	0.8802	0.8746	0.8730	135
weighted avg	0.8804	0.8741	0.8729	135

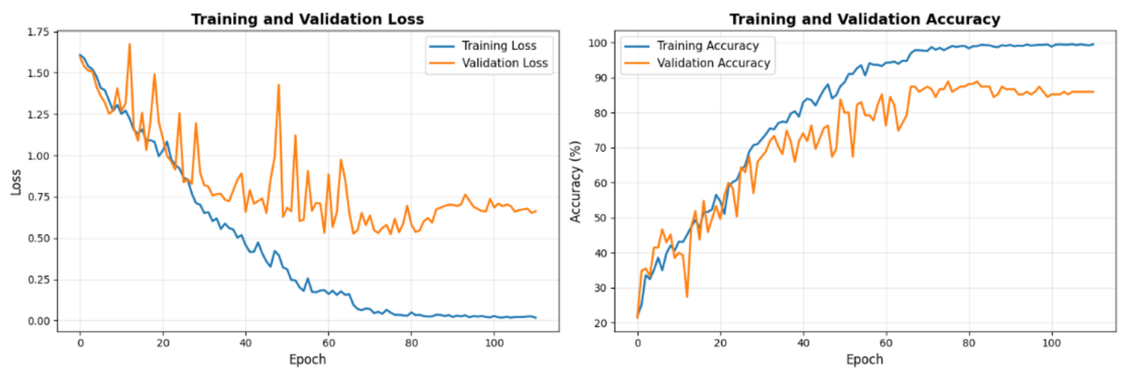
PER-CLASS ACCURACY
Emotion 1: 89.29% (28 samples)
Emotion 2: 85.19% (27 samples)
Emotion 3: 74.07% (27 samples)
Emotion 4: 96.15% (26 samples)
Emotion 5: 92.59% (27 samples)

```
Best parameters: {'max_depth': 25, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 1
50}
Best CV accuracy: 0.7556 (75.56%)
Test set accuracy: 0.7778 (77.78%)
```



- Training curves show overfitting
- Despite this, the test accuracy remains stable (~88%) across multiple random seeds.
- The consistent test performance suggests the model has reached the dataset's generalization plateau.
- The gap between training and validation performance likely reflects data noise, not model instability.






```
Training Summary:
Total epochs: 111
Best validation accuracy: 88.89%
Final training accuracy: 99.46%
Final validation accuracy: 85.93%
```

The CNN-LSTM model significantly outperformed the Random Forest
Key findings include:

- CNN-LSTM Accuracy: 87.41%
- Random Forest Accuracy: 77.78%
- Augmentation greatly improved generalization
- Highest confusion occurred between sad and neutral

Performance demonstrates the advantage of spectrogram-based deep learning for emotion recognition.

Github

Ku.chen@ufl.edu haichenfan@ufl.edu
zixuan.liu@ufl.edu