SPEECH EMOTION RECOGNITION

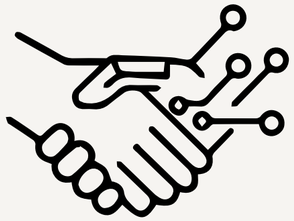
Deep Learning on the RAVDESS Dataset

Presented By
CHRIS DOMINIC ESTREBA

PROJECT **OVERVIEW**

Goal

Why it matters



Proposed Solution

A Convolutional Neural Network (CNN) wrapped in an interactive Streamlit web app

Streamlit

DATASET **OVERVIEW**

Dataset

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Actors

24 professional actors (12 male, 12 female)

Controlled Environment

Actors spoke two standard sentences *("Kids are talking...", "Dogs are sitting...")* to isolate emotion from context.

The 8 Emotions

- Neutral
- Calm
- Happy
- Sad
- Angry
- Fearful,
- Disgust
- Surprised

DATA
PRE-0PROCESSING

Data Strategy

- **Reproducibility**: Global Random Seed = 42
- **Stratification:** Splits balanced by emotion to ensure fair evaluation

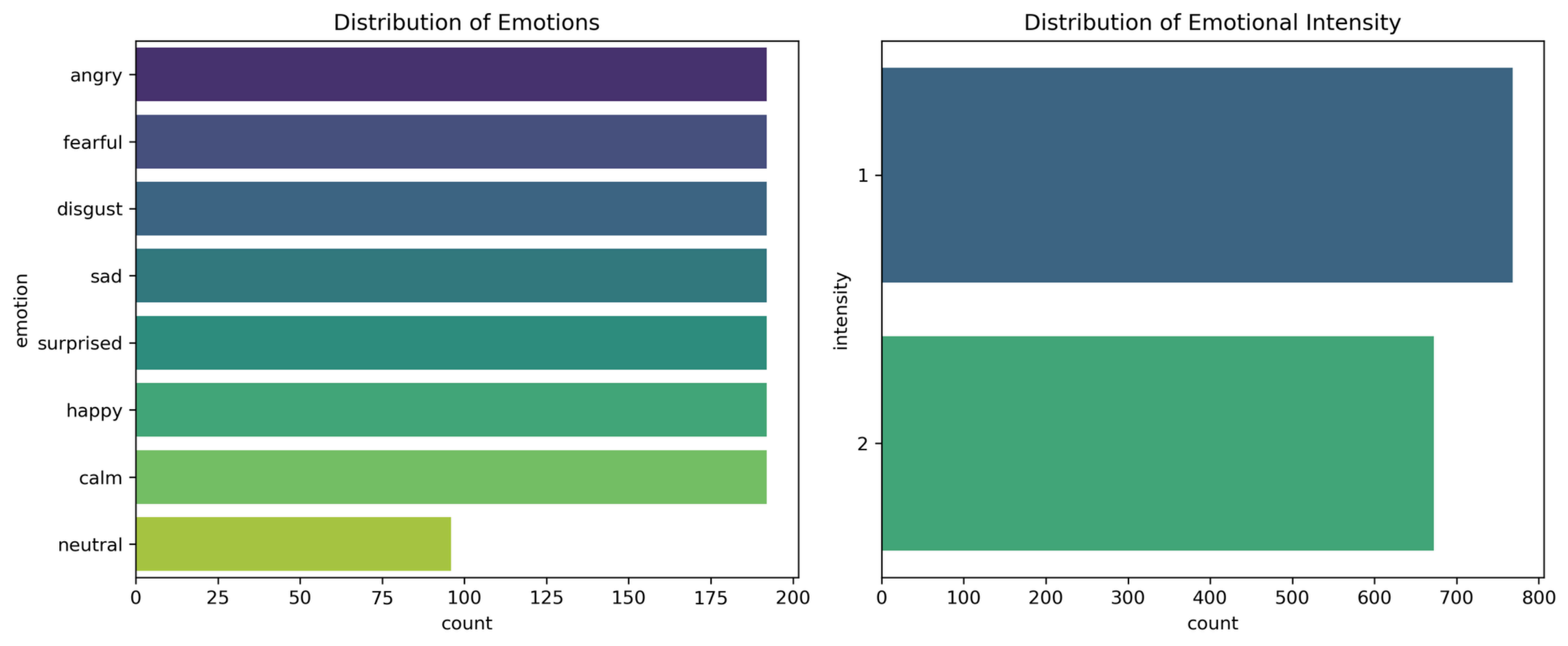
Data Splits

- **Deep Learning (CNN):** 70% Train / 15% Validation / 15% Test
- **Classical ML:** 70% Train / 30% Test

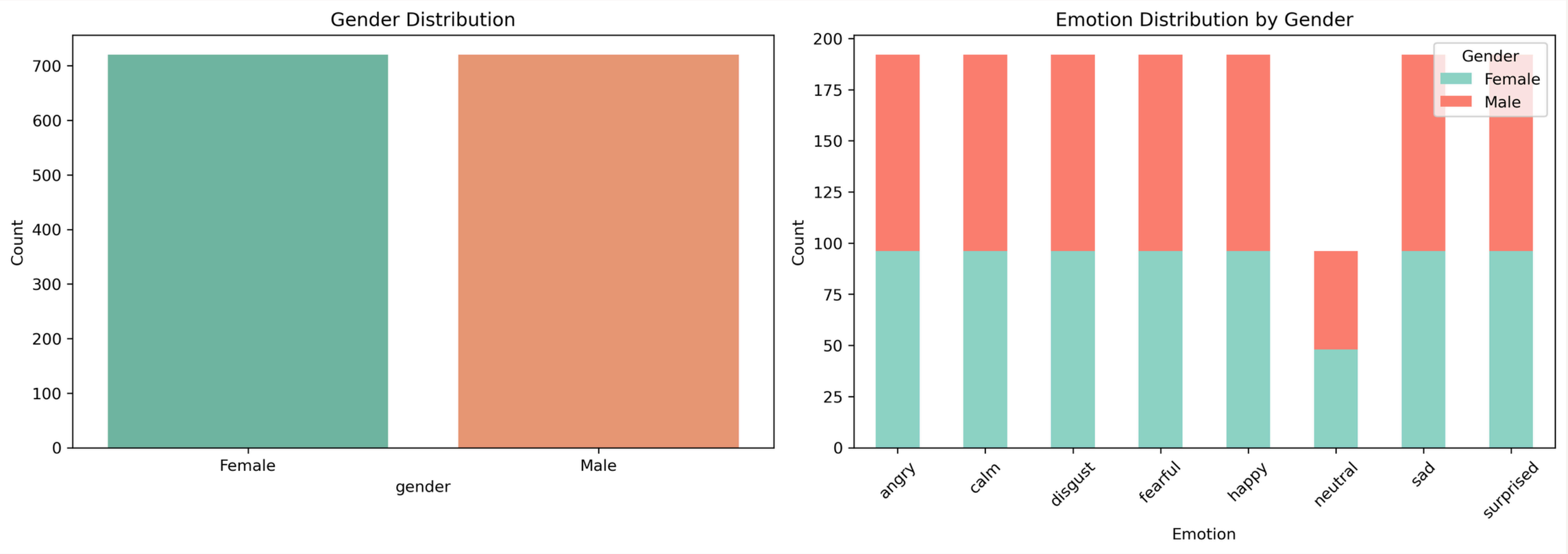
Pre-processing

- **Classical ML:** `StandardScaler` for feature normalization
- **Deep Learning:** Log-Mel Spectrogram conversion (Audio -> Image).

DATASET
OVERVIEW



DATASET **OVERVIEW**



METHODOLOGY
EVOLUTION OF OUR APPROACH

phase 1

- Raw Audio Preprocessing
  - trimming leading/trailing silence
  - **top_db=25**
- All clips trimmed to **~4.08s** (89,996 samples)
- **90k features** = audio samples + metadata

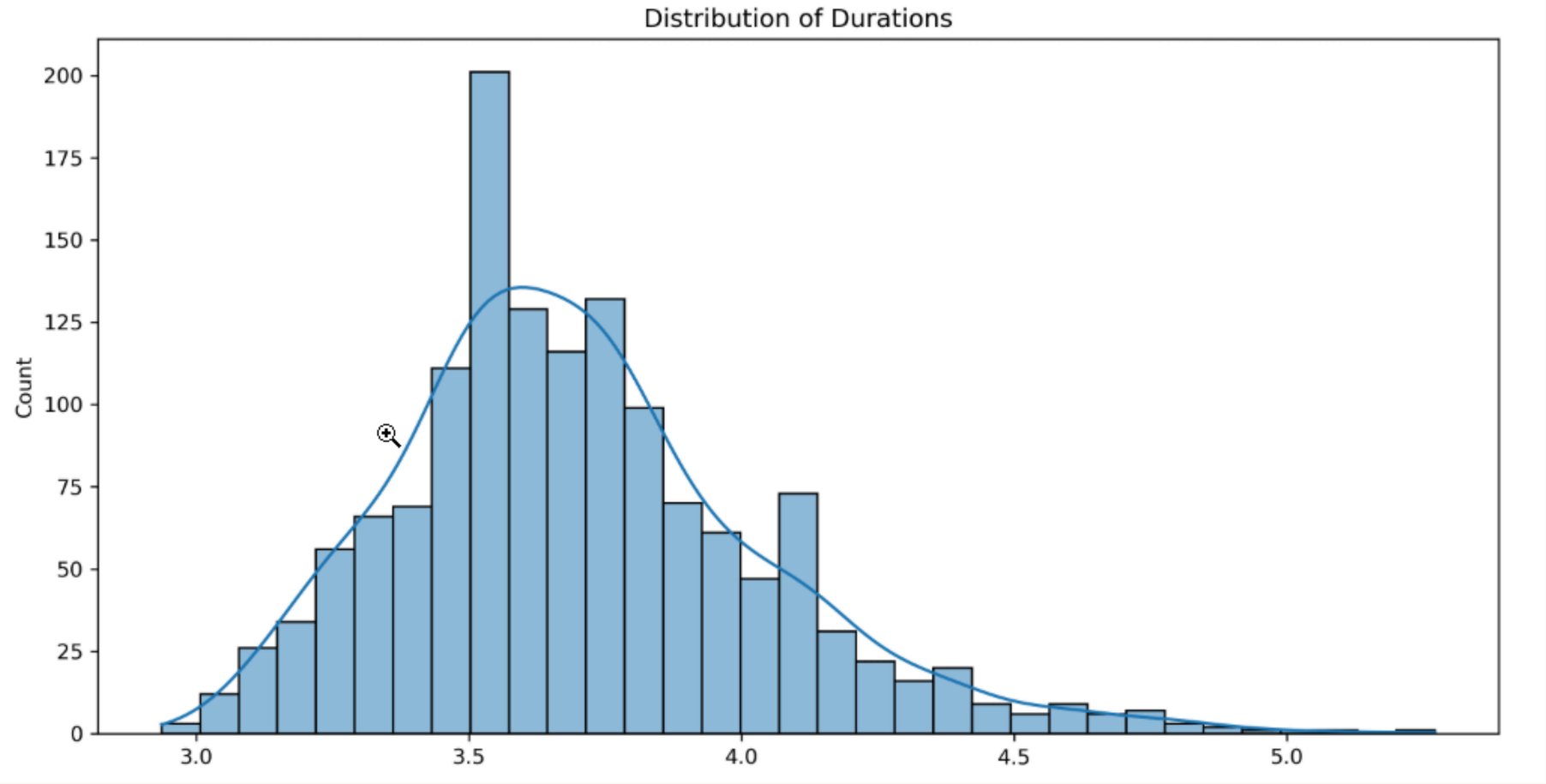
phase 2

- Reduced dimensionality
- Extracted **MFCCs**
  - Mel-frequency cepstral coefficients
- Multi-Layer Perceptron

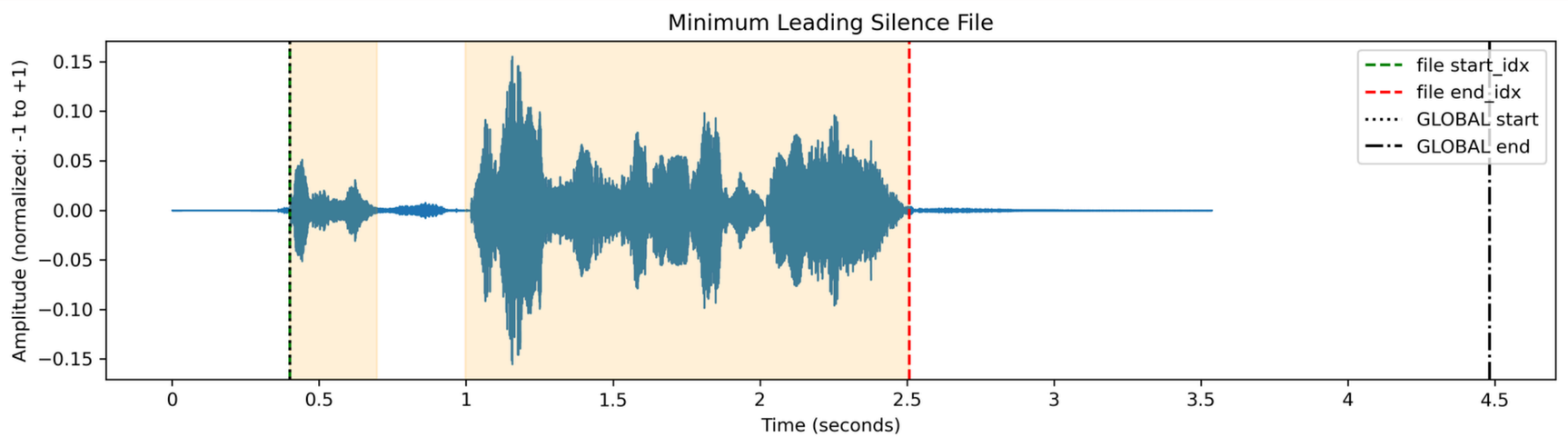
phase 3

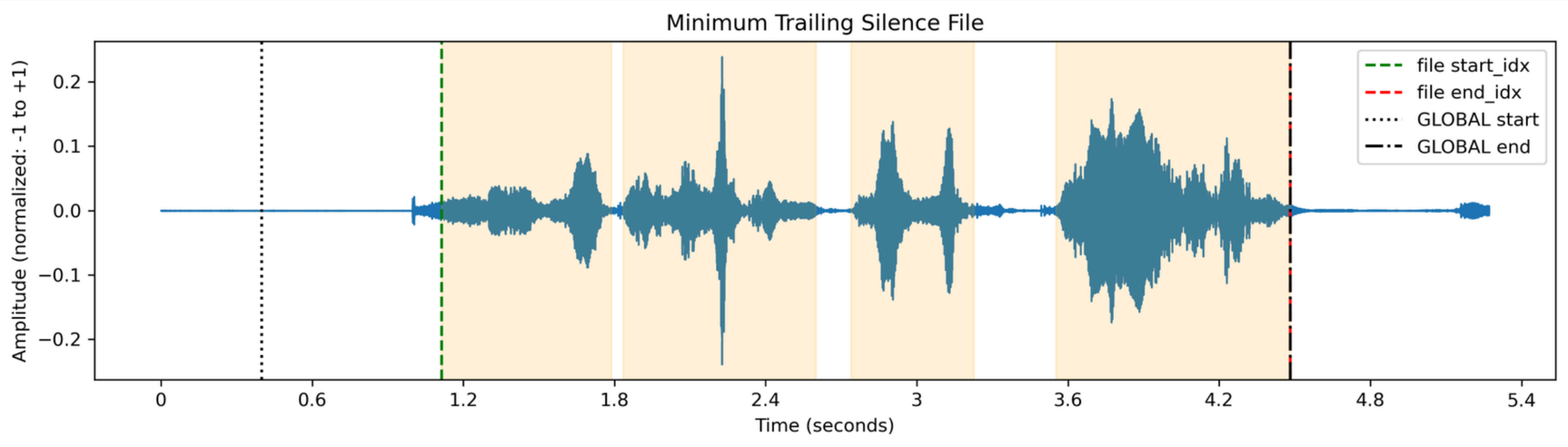
- Treat sound as an image to keep temporal patterns
- Converted audio to **Mel Spectrograms**
- Convolutional Neural Network (CNN)

THE "RAW" ATTEMPT
PHASE 1



THE "RAW" ATTEMPT
PHASE 1





THE "RAW" ATTEMPT
PHASE 1

	modality	vocal_channel	intensity	statement	repetition	actor	emotion	gender	loudness_db	y_0	...	y_89986	y_89987	y_89988	y_89989	y_89990	y_89991	y_89992	y_89993	y_89994	y_89995
0	3	1	1	2	1	16	angry	Female	-64.729790	0.000000e+00	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	3	1	1	2	2	16	fearful	Female	-65.943779	2.051093e-08	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	3	1	2	1	2	16	fearful	Female	-51.392780	-8.404510e-06	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	3	1	2	1	1	16	angry	Female	-59.114307	0.000000e+00	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	3	1	1	1	1	16	disgust	Female	-69.102135	0.000000e+00	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1435	3	1	2	2	2	8	happy	Female	-57.352631	8.016690e-05	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1436	3	1	1	1	2	8	happy	Female	-64.781517	4.014514e-07	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1437	3	1	2	1	1	8	calm	Female	-67.697044	-4.947469e-05	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1438	3	1	1	2	1	8	calm	Female	-66.581841	-9.008037e-05	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1439	3	1	1	2	2	8	neutral	Female	-65.417435	1.524629e-05	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

1440 rows × 90005 columns

- **modality** (3) Audio-visual indicator (constant for this dataset)
- **vocal_channel** (1) Speech indicator (constant for this dataset)
- **intensity** (1 or 2) Normal or strong emotional intensity
- **statement** (1 or 2) Which statement was spoken
- **repetition** (1 or 2) First or second repetition
- **actor** (1–24) Actor ID (odd=male, even=female)
- **gender** (Male/Female) Derived from actor ID
- **loudness_db** Average loudness in decibels

MODEL ARCHITECTURE
TRADITIONAL ML

The "Overfitting Trap":

Models achieved **99-100% Training Accuracy,**
but stalled at **~33-35% Test Accuracy**

PCA/LDA Result: Dimensionality reduction **did not solve** the problem

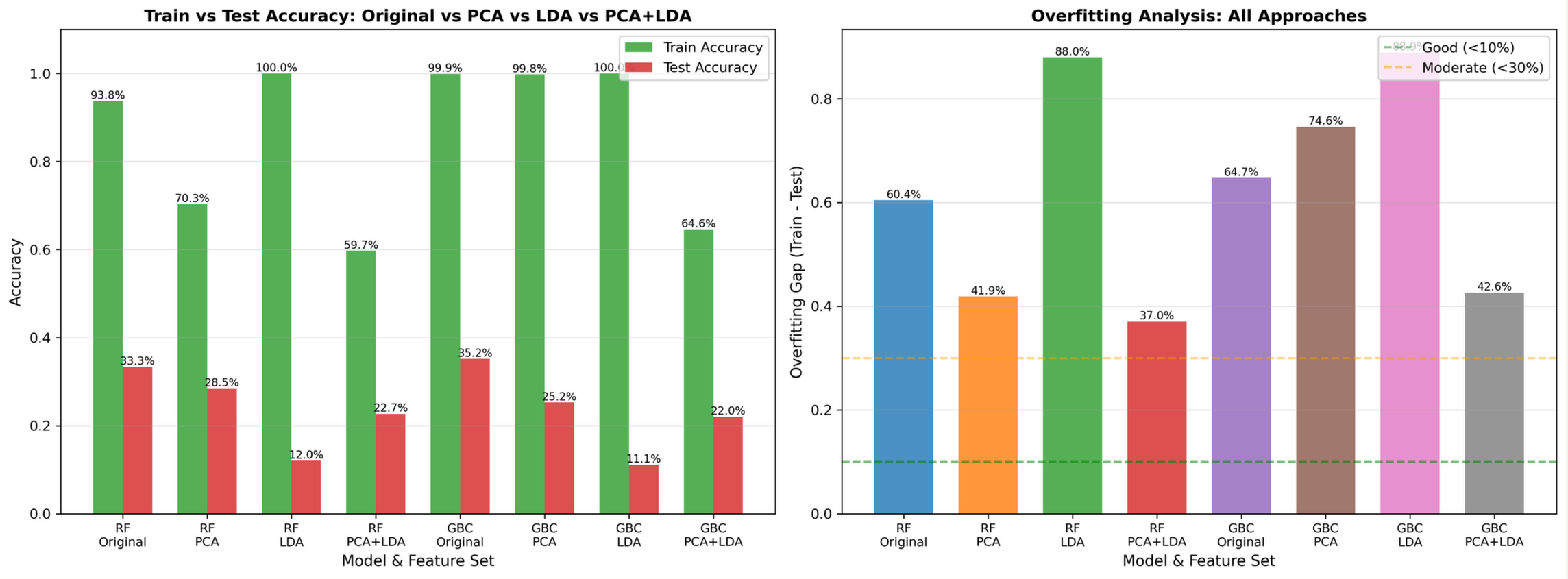
- Even with 30 trials and optimized params
  - accuracy remained low (~22-23%),
  - proving that linear separation of this **reduced feature space was insufficient**.

Key Takeaway:

- **Curse of Dimensionality**
- Classical models **memorized noise**
- **failed to capture non-linear** emotional invariants in the reduced space.
- **90k raw features**. Dimensionality reduction (PCA/LDA) was explored but complex.

	Train Accuracy	Test Accuracy
Random Forest (Original)	93.7%	33.3%
Random Forest (PCA)	70.37%	28.5%
Random Forest (LDA)	100%	12%
Random Forest (PCA+LDA)	59.7%	22.7%
GBC (Original)	99%	35%
GBC (PCA)	99.8%	25.2%
GBC (LDA)	100%	11%
GBC (PCA+LDA)	64.6%	22%

RESULTS & KEY **FINDINGS**



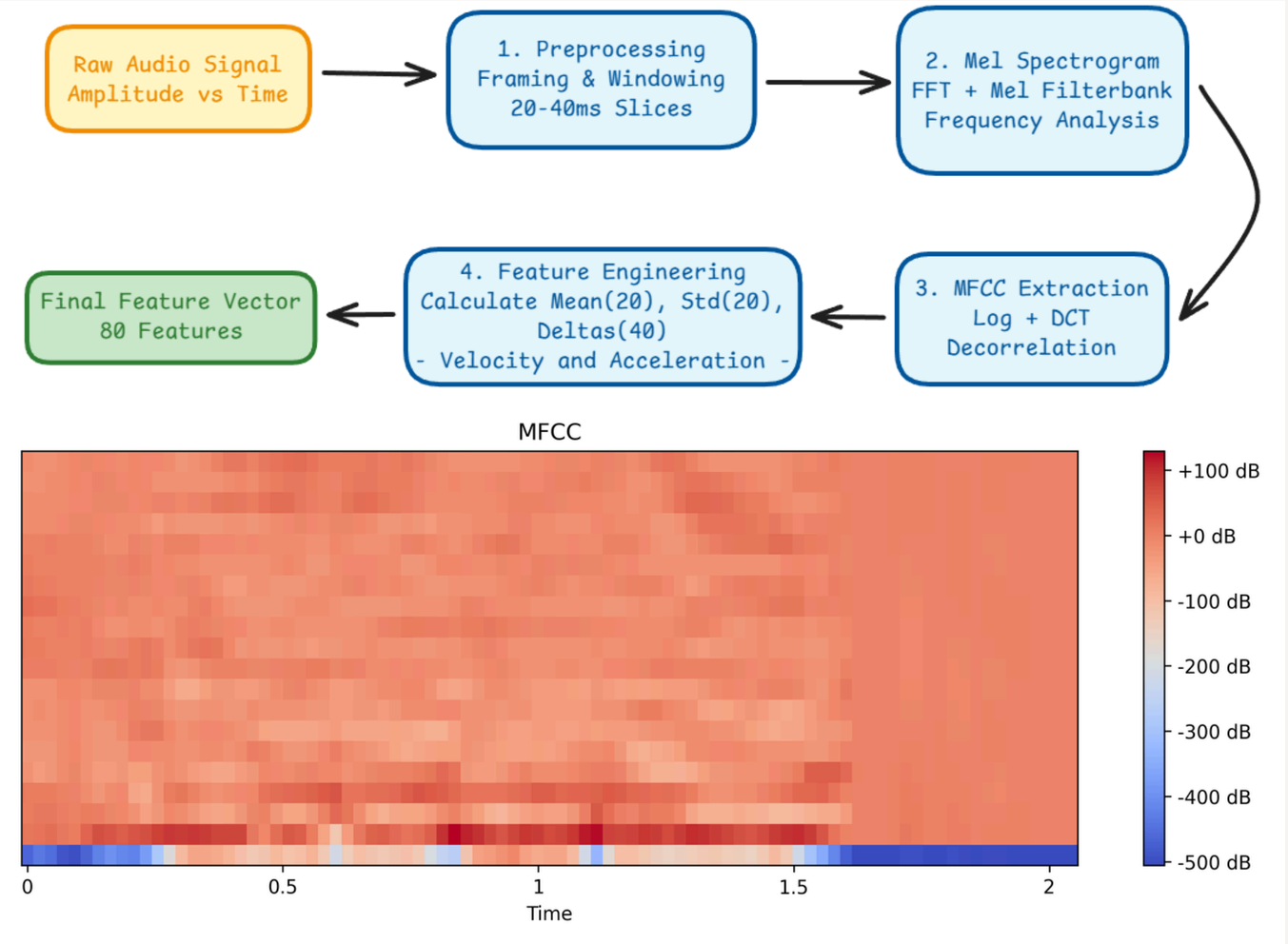
METHODOLOGY
PHASE 2

Curse of Dimensionality

Solution: Extracted **MFCCs**

Reduction: Compressed 90k raw features down to ~80 statistical features

Model: Trained a Multilayer Perceptron (MLP). Performance improved, but we lost temporal detail.



METHODOLOGY
PHASE 2

	file_path	emotion	mfcc1_mean	mfcc1_std	mfcc2_mean	mfcc2_std	mfcc3_mean	mfcc3_std	mfcc4_mean	mfcc4_std
0	/home/nico/ds_workspace/projects/RAVDESS/data/...	neutral	-697.792603	183.030441	54.890038	72.168480	0.663467	19.195799	12.435785	20.930756
1	/home/nico/ds_workspace/projects/RAVDESS/data/...	neutral	-692.855774	185.050293	55.363899	66.308495	-1.548319	19.290407	16.038307	19.345299
2	/home/nico/ds_workspace/projects/RAVDESS/data/...	neutral	-691.587891	190.336121	58.024662	72.184830	0.159465	21.651524	13.624649	19.525526
3	/home/nico/ds_workspace/projects/RAVDESS/data/...	neutral	-685.105469	184.565063	55.879421	66.488159	2.783262	20.100769	13.252023	20.778818
4	/home/nico/ds_workspace/projects/RAVDESS/data/...	calm	-727.104370	182.821884	62.355034	68.404228	3.121181	22.141096	15.064669	20.880312
5	/home/nico/ds_workspace/projects/RAVDESS/data/...	calm	-707.358215	169.380035	66.736458	73.044151	2.253490	22.743551	11.169915	19.891697
6	/home/nico/ds_workspace/projects/RAVDESS/data/...	calm	-697.166138	195.661469	65.108200	77.897346	0.930369	25.643011	13.633629	20.614908
7	/home/nico/ds_workspace/projects/RAVDESS/data/...	calm	-698.637695	196.158691	68.698586	75.416084	1.100368	22.913017	13.685941	19.570847
8	/home/nico/ds_workspace/projects/RAVDESS/data/...	calm	-734.125610	190.543060	70.532913	71.599319	4.225180	20.327414	13.866501	21.240456
9	/home/nico/ds_workspace/projects/RAVDESS/data/...	calm	-697.822327	173.384033	67.339592	72.696190	-0.449553	21.242195	11.884349	18.541895

10 rows × 82 columns

METHODOLOGY
PHASE 3

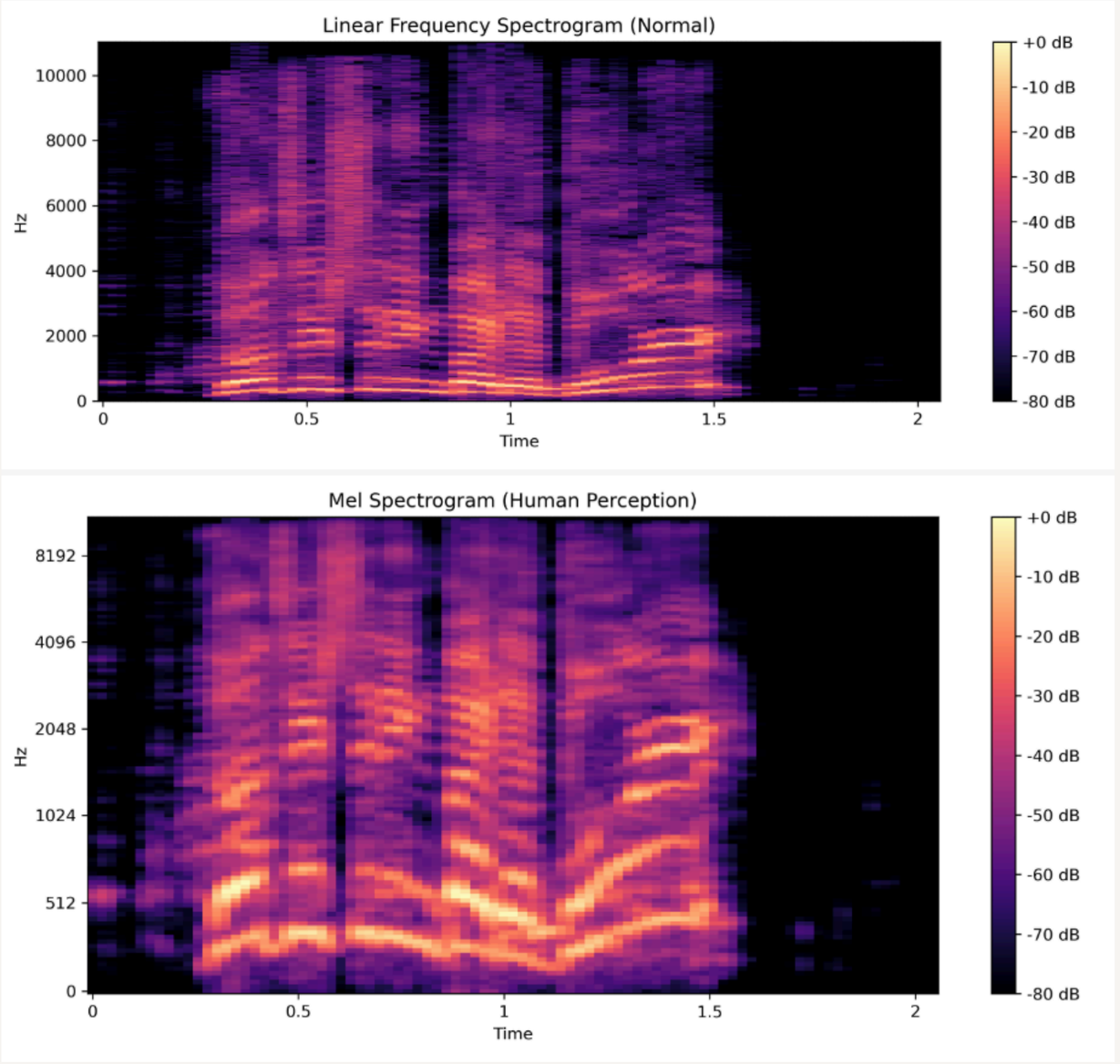
Why Mel Spectogram?

Idea: Treat sound as an image to keep temporal patterns

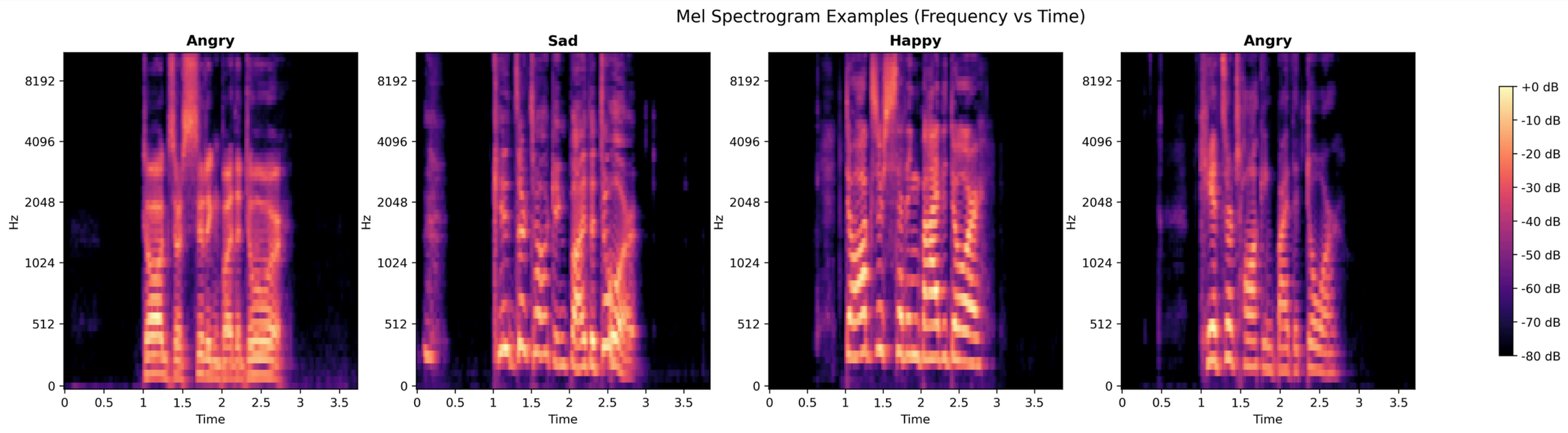
Input: Converted audio to **Mel Spectrograms**

Linear: Shows all frequencies equally. Good for physics, bad for perception

Mel: Stretches/compress frequencies to match how **humans hear**
(more detail in low frequencies, less in high)



METHODOLOGY
PHASE 3



MODEL ARCHITECTURE
DEEP LEARNING

MLP

- **Input:** MFCCs (80 statistical features)
- **Tool:** PyCaret
- **Result: ~62%** accuracy
  - Fixed dimensionality, but lost temporal detail

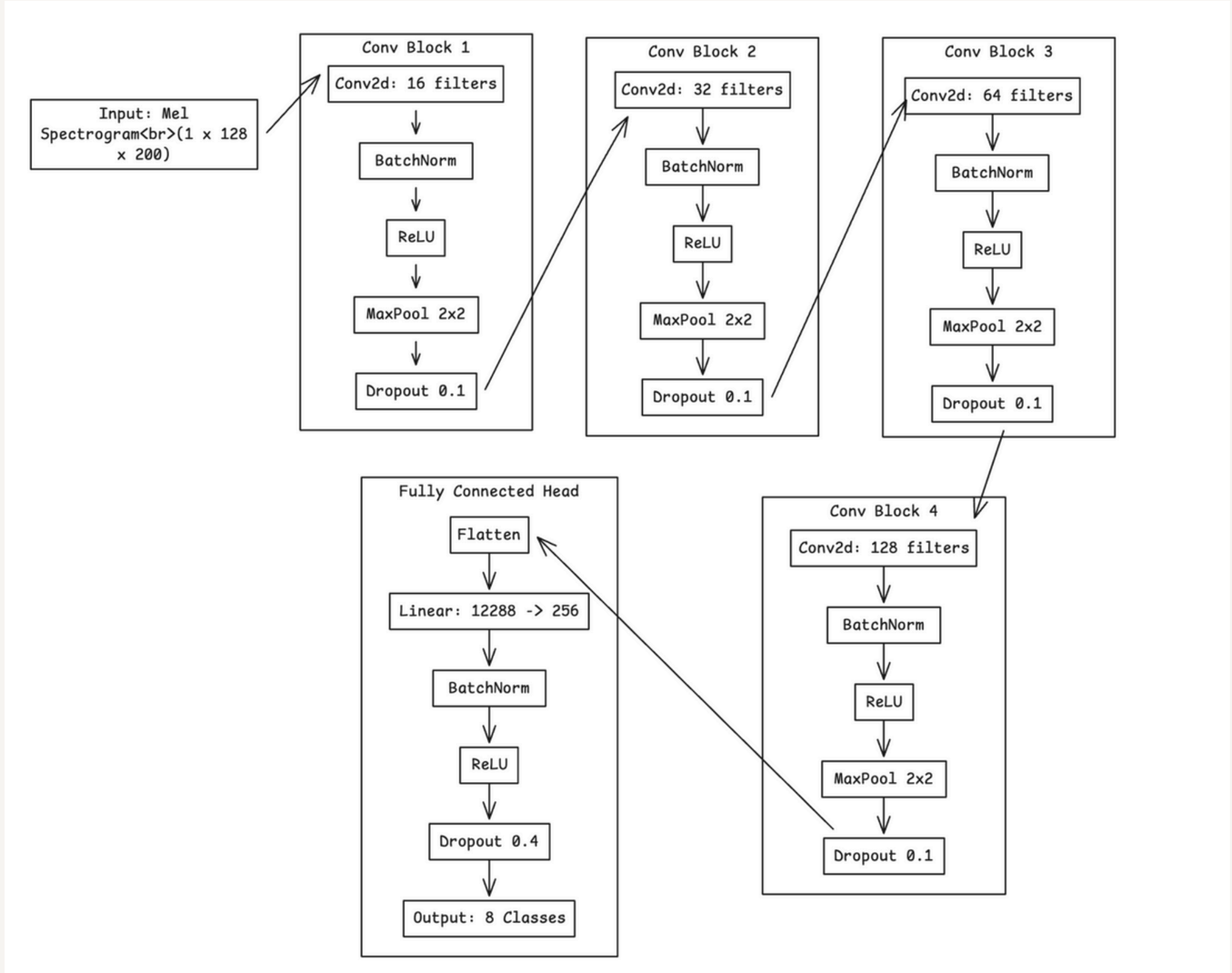
CNN

- **Architecture:** 4-Block Deep CNN (up to 128 filters) with Dropout (0.4) for regularization.
- **Input:** Normalized Mel Spectrograms (200 time steps).
- **Performance: ~69%** Validation Accuracy (Test: 68%).

MODEL ARCHITECTURE
DEEP LEARNING

CNN

- **Architecture:**
  - Each with Conv2d
  - (kernels: 16–>32–>64–>128), BatchNorm, ReLU, MaxPool(2x2), and Dropout(0.1).
- **Input:** Normalized Mel Spectrograms (200 time steps).
- **Performance: ~69%** Validation Accuracy (Test: 68%).

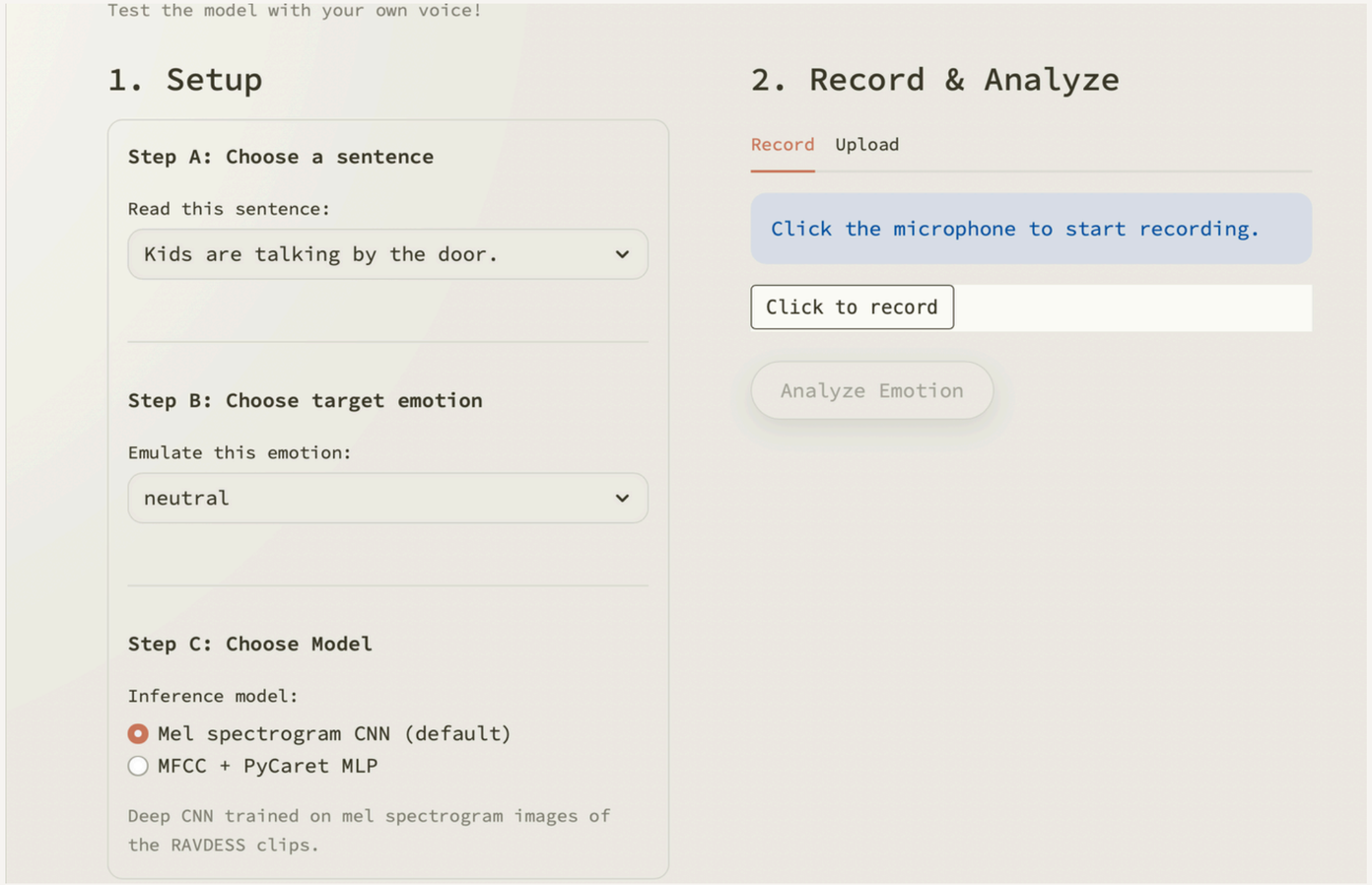


INTERACTIVE DEMO **STREAMLIT**

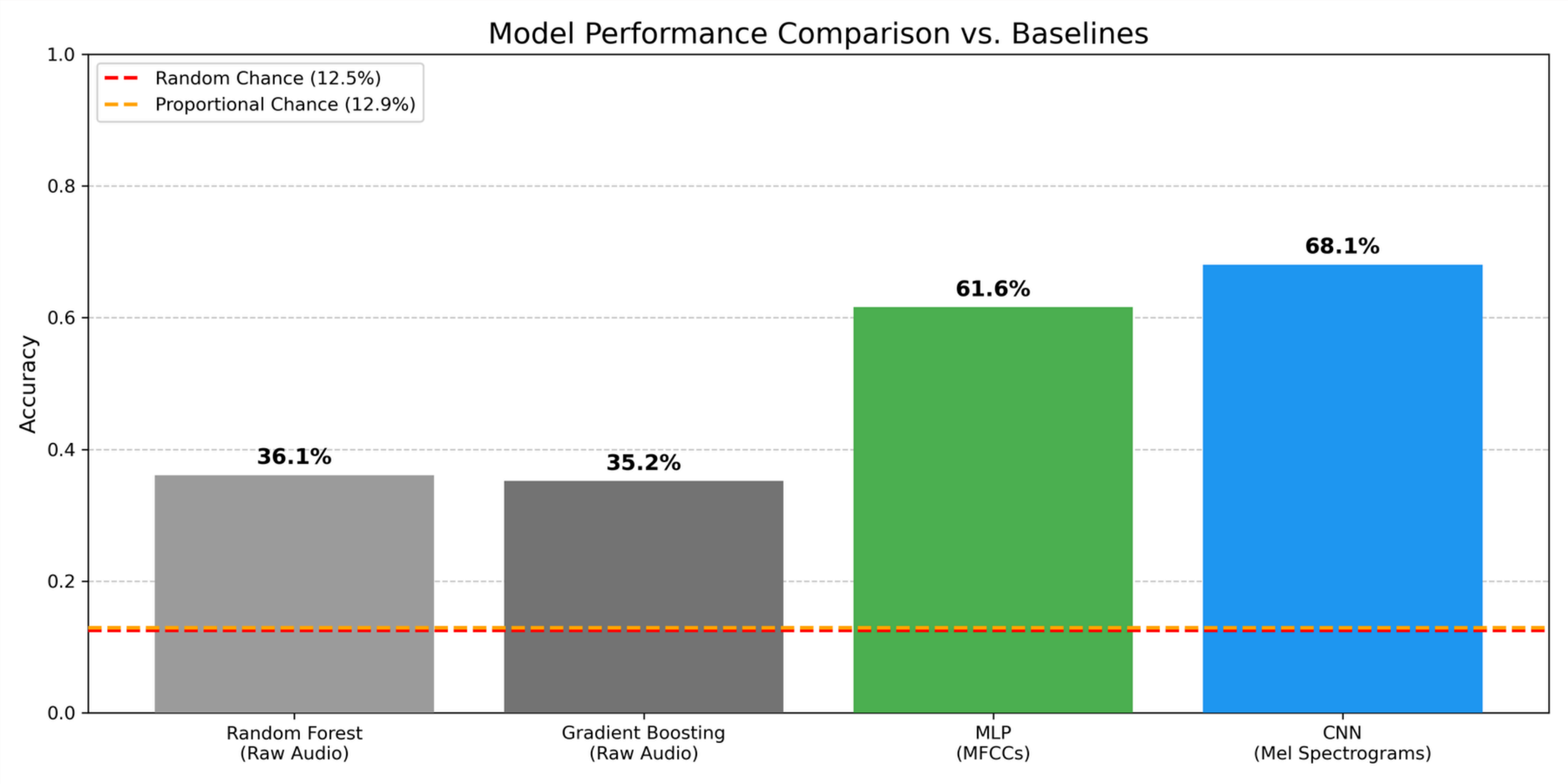
Tech Stack: Python + Streamlit + PyTorch.

Features:

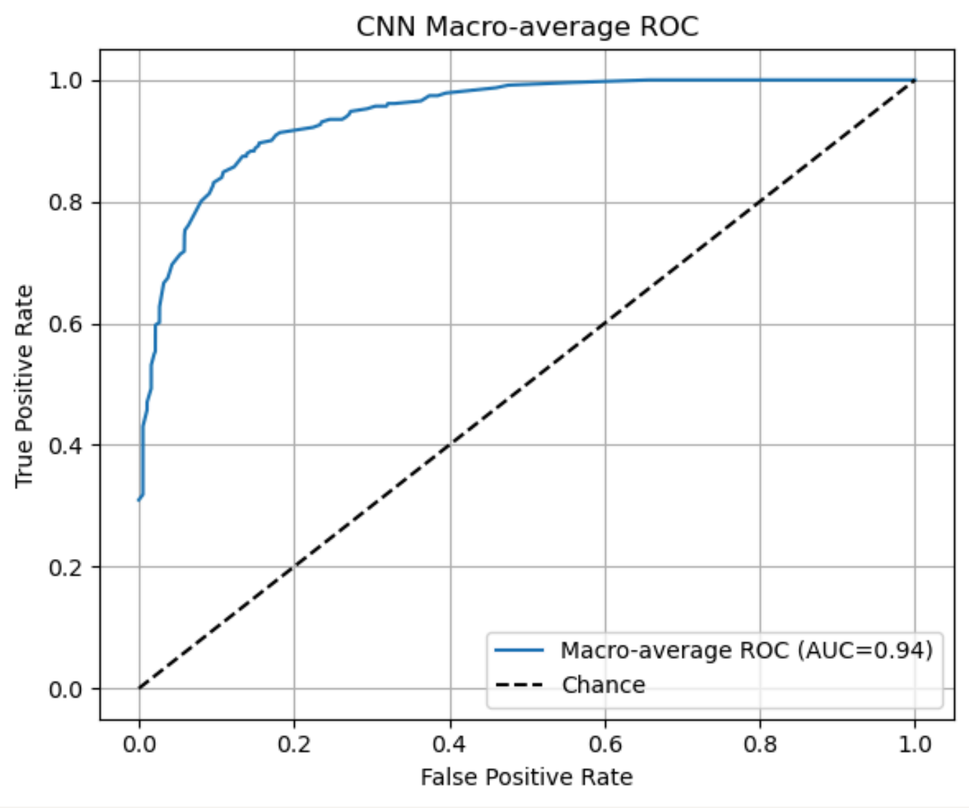
- **Live Recording:** Record directly in-browser.
- **Real-time Inference:** Instant feedback.
- **Visualization:** Probability bars showing model confidence.

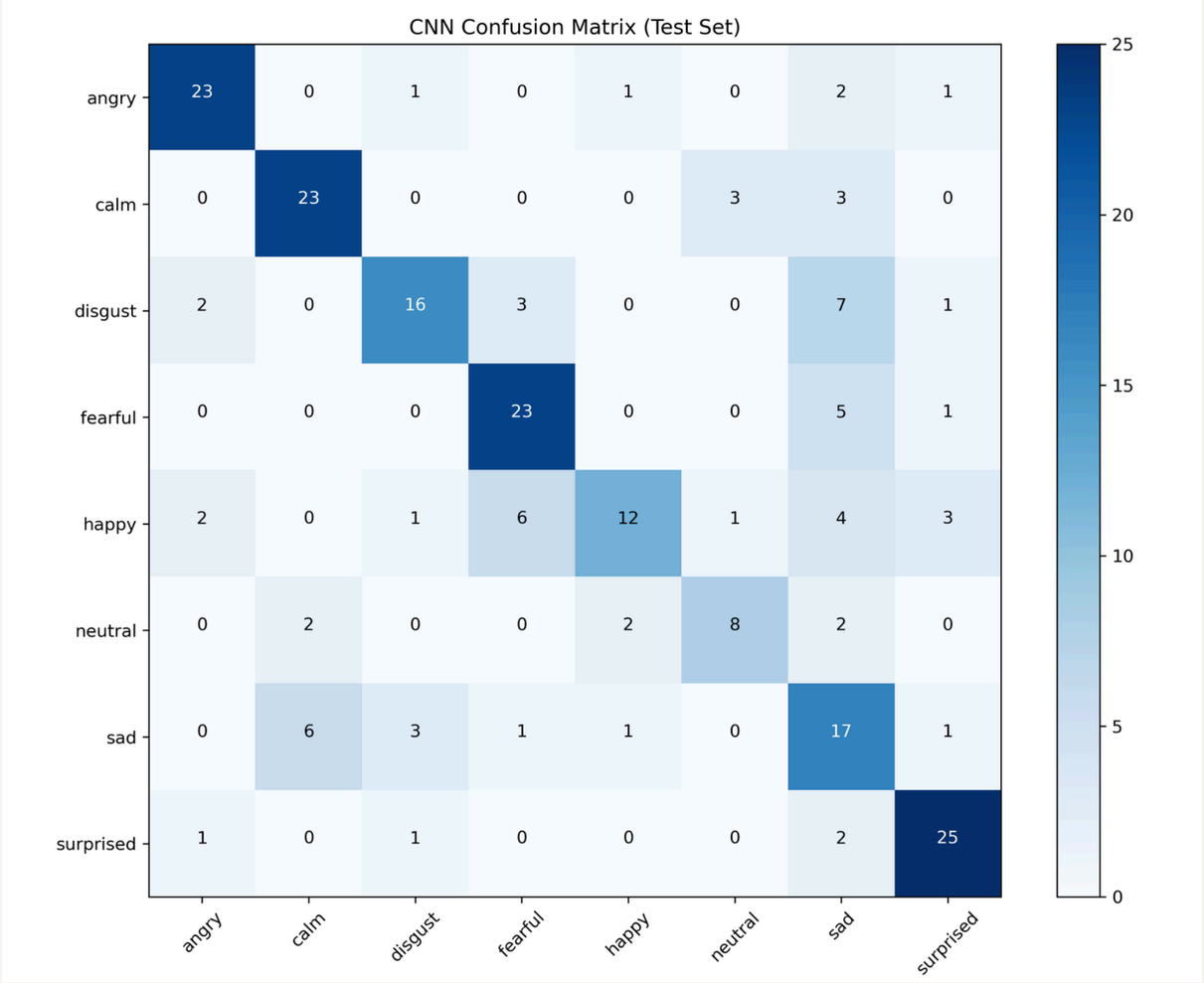


RESULTS & KEY **FINDINGS**



RESULTS & KEY **FINDINGS**





RESULTS & KEY **FINDINGS**

Model Comparison:

Classical ML (Baselines)

- **Random Forest** (Optuna Tuned): ~**23%** (on PCA+LDA features) / **36%** (on Raw Audio)
- **Gradient Boosting** (Optuna Tuned): ~**22%** (on PCA+LDA features)

DeepLearning (Baselines

- Improved Mel-CNN: **~69.0%**

Baselines: .

- **Random Chance**: 12.5% (1/8 classes)
- **Proportional Chance**: ~12.9% (Weighted by class distribution)

Key Takeaway

All models beat the random baseline, but only the CNN provides strong predictive power (5x better than random), whereas traditional ML struggles to break away from the "noise floor" of ~36%.

RESULTS & KEY **FINDINGS**

Challenges:

Overfitting: 1,440 clips is a small dataset for Deep Learning.

Similar Emotions: "Calm" and "Neutral" are hard to distinguish.

RECOMMENDATIONS

Data

- Limited dataset
  - only 1440 samples
- Expand dataset with **CREMA-D/TESS** and apply **rigorous augmentation** (Noise/Pitch Shift) to combat overfitting.

Model

- Adopt Transfer Learning (Wav2Vec 2.0) and Hybrid architectures (CRNN) to **capture temporal dynamics**.

Deployment

- Implement **real-time noise reduction** and a user feedback loop for continuous model improvement.