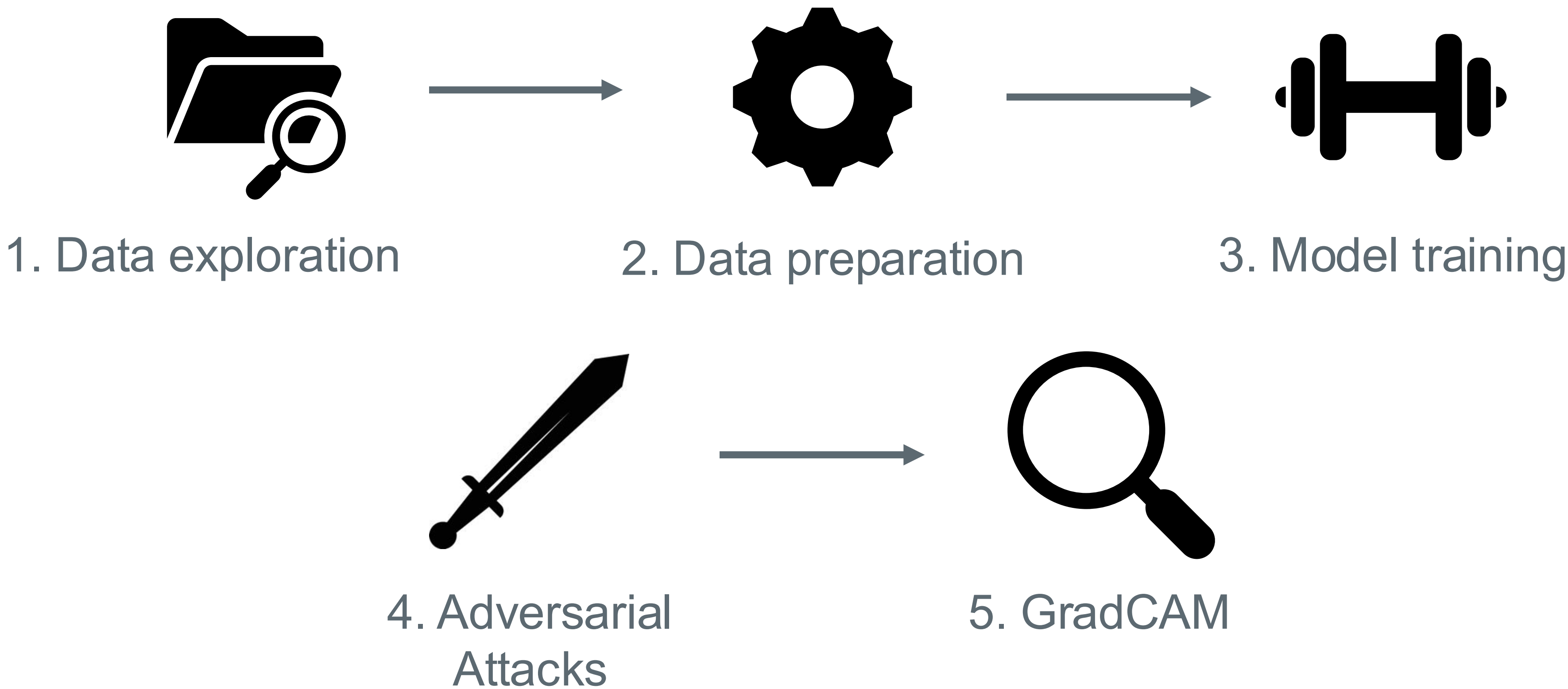
Duale Hochschule
Baden-Württemberg

December 8, 2025

Robustness of a CNN for music genre classification

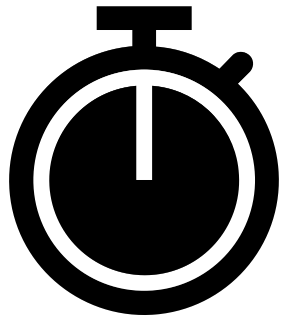
Erik von Heyden (8720832)

Methodology



GTZAN Dataset

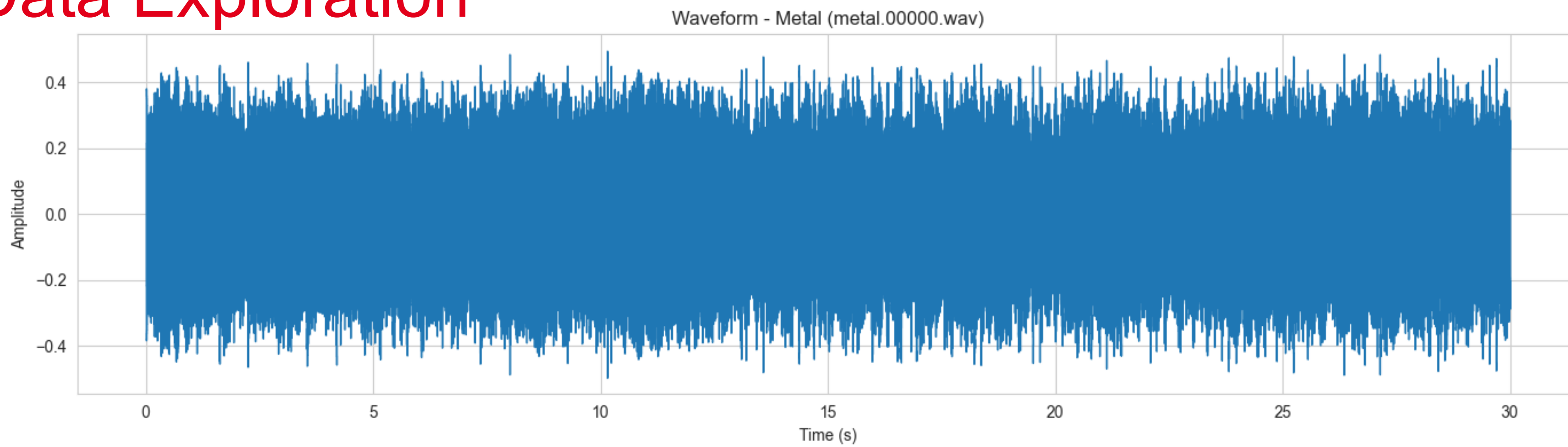
1000 songs
in 10 genres
(100 per genre)

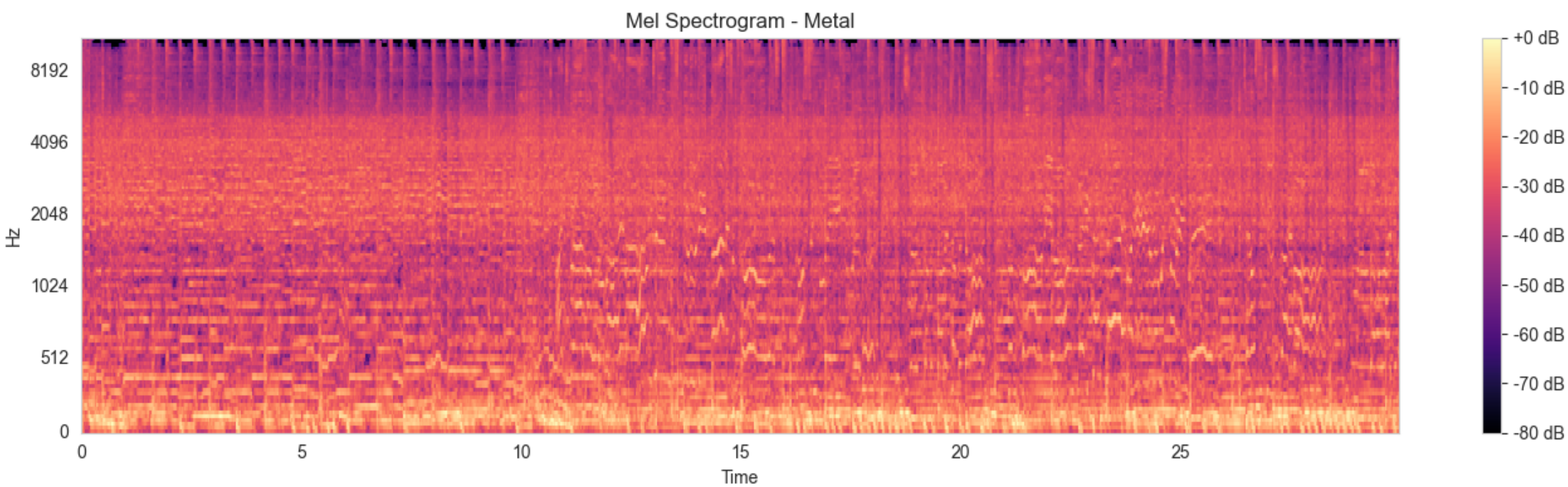


30 sec
as .wav

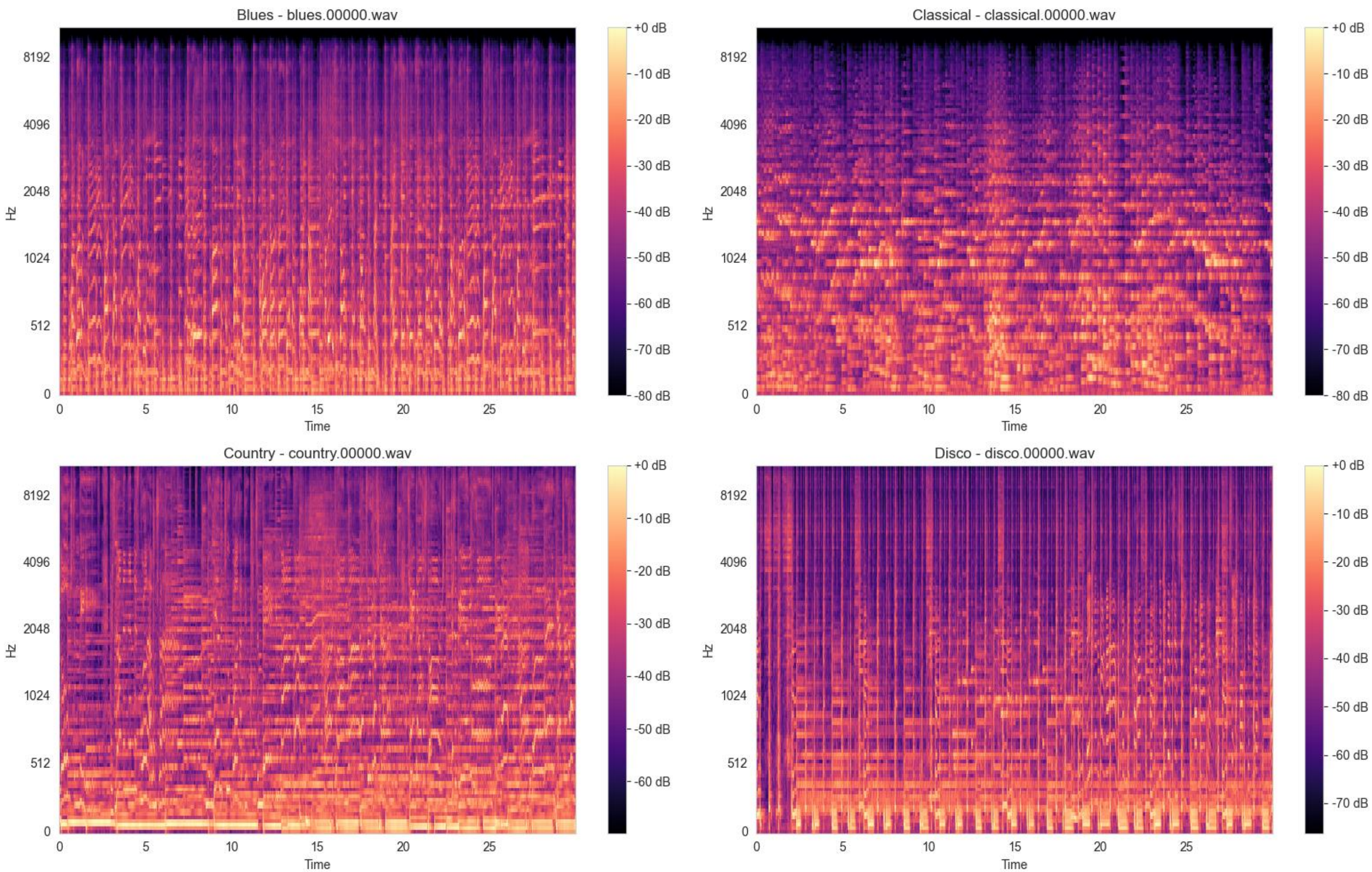
Mel spectograms,
feature tables

Data Exploration





Data Exploration



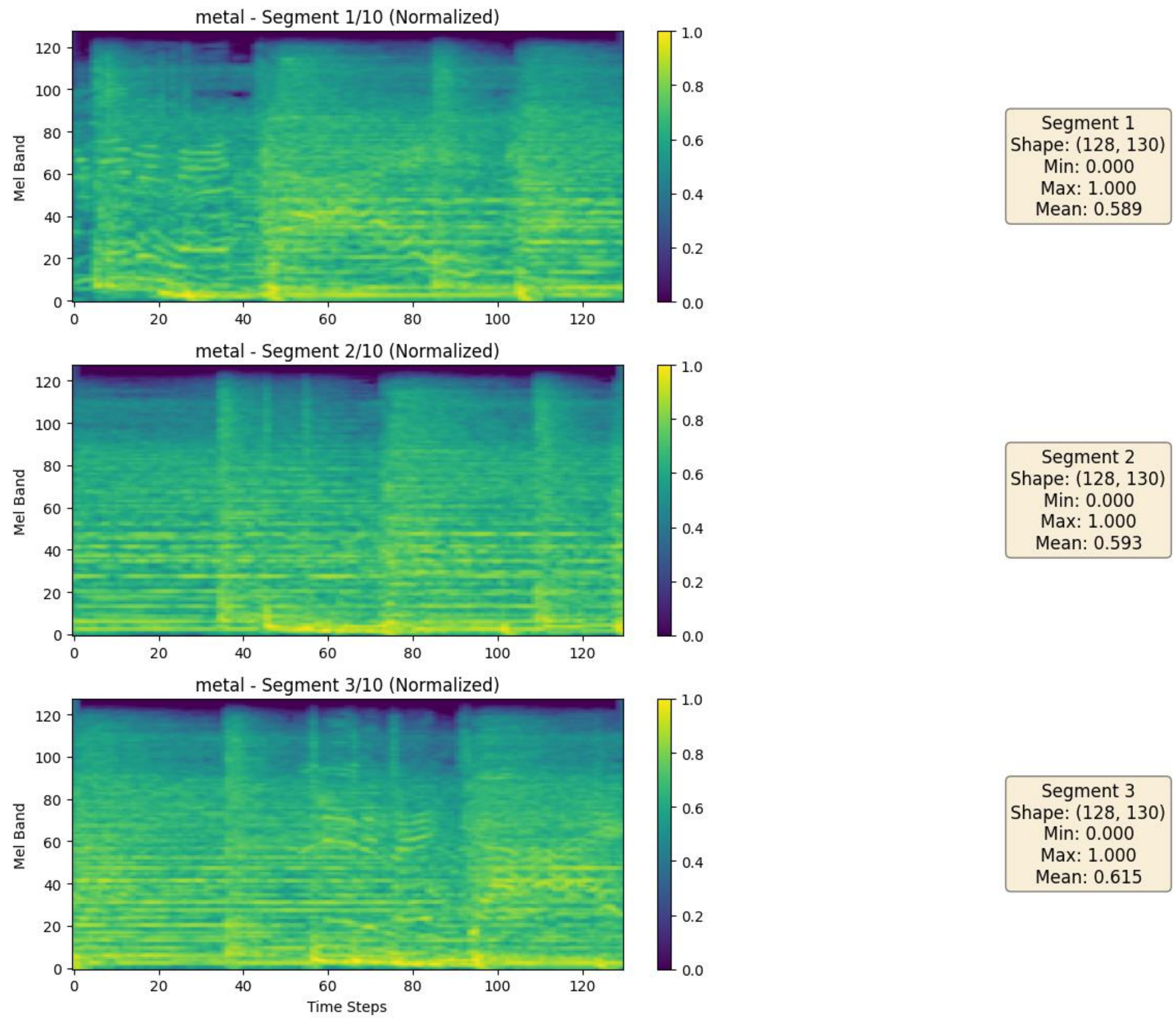
Preprocessing

- 70% Training (~700 Songs)
- 15% Validation (~150 Songs)
- 15% Test (~150 Songs)

Split into 10 segments á 3 sec

- 70% Training (~7000 segments)
- 15% Validation (~1500 segments)
- 15% Test (~1500 segments)

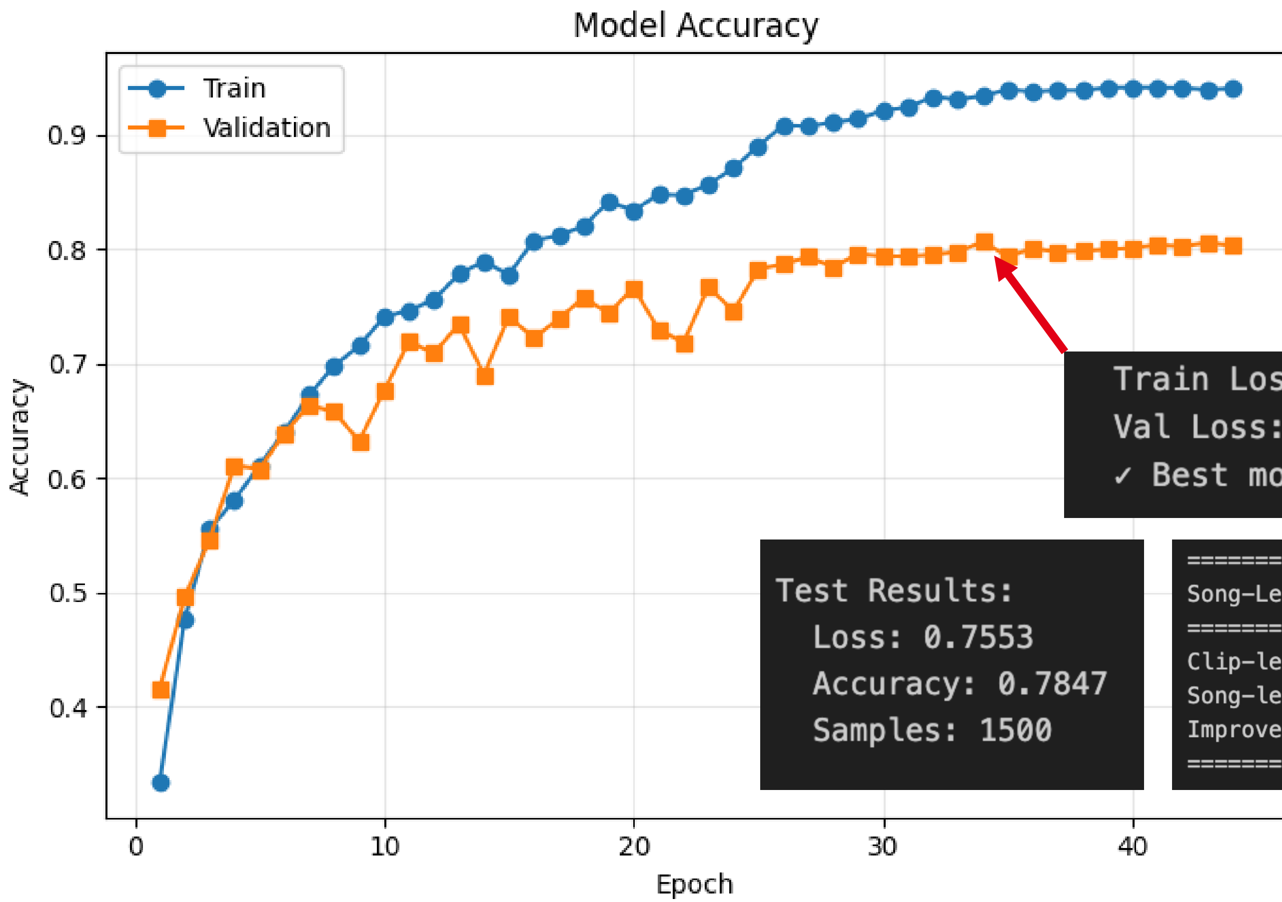
librosa


metal - Segment 1/10 (Normalized)
Mel Band
0
20
40
60
80
100
120
1.0
0.8
0.6
0.4
0.2
0.0
Segment 1
Shape: (128, 130)
Min: 0.000
Max: 1.000
Mean: 0.589
metal - Segment 2/10 (Normalized)
Mel Band
Segment 2
Shape: (128, 130)
Min: 0.000
Max: 1.000
Mean: 0.593
metal - Segment 3/10 (Normalized)
Mel Band
Time Steps
Segment 3
Shape: (128, 130)
Min: 0.000
Max: 1.000
Mean: 0.615

CNN Architecture

- Input: Mel Spectrogram 128 × 130 × 1
- Convolutional Blocks: 2x (Conv→ BatchNorm) → MaxPool → Dropout
- Progressive filters: 32 → 64 → 128 → 256
- Classification: Flatten → Dense(512) → Dense(10)
- 1,4M parameters

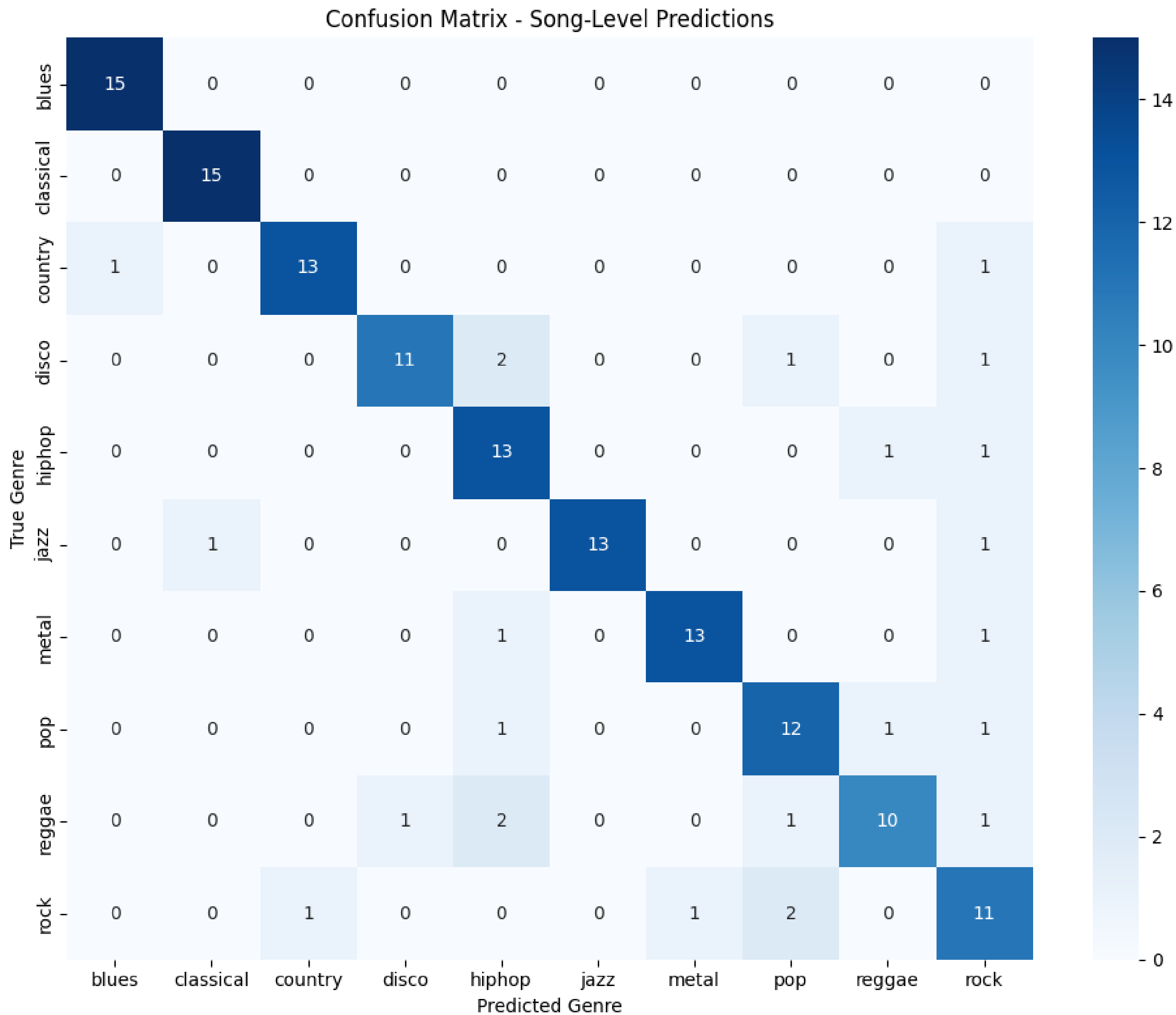
Training Results




```
Train Loss: 0.1983, Train Acc: 0.9341
Val Loss: 0.8362, Val Acc: 0.8064
✓ Best model saved (val_acc: 0.8064)
```

```
Test Results:
  Loss: 0.7553
  Accuracy: 0.7847
  Samples: 1500
```

```
============================================================
Song-Level Evaluation (Soft Voting)
============================================================
Clip-level accuracy: 0.7847 (1500 clips)
Song-level accuracy: 0.8400 (150 songs)
Improvement: +5.53%
============================================================
```


Confusion Matrix - Song-Level Predictions
True Genre
Predicted Genre
blues
classical
country
disco
hiphop
jazz
metal
pop
reggae
rock
blues: 15 0 0 0 0 0 0 0 0 0
classical: 0 15 0 0 0 0 0 0 0 0
country: 1 0 13 0 0 0 0 0 0 1
disco: 0 0 0 11 2 0 0 1 0 1
hiphop: 0 0 0 0 13 0 0 0 1 1
jazz: 0 1 0 0 0 13 0 0 0 1
metal: 0 0 0 0 1 0 13 0 0 1
pop: 0 0 0 0 1 0 0 12 1 1
reggae: 0 0 0 1 2 0 0 1 10 1
rock: 0 0 1 0 0 0 1 2 0 11
14
12
10
8
6
4
2
0

Adversarial Attacks

FGSM (Fast Gradient Sign Method)

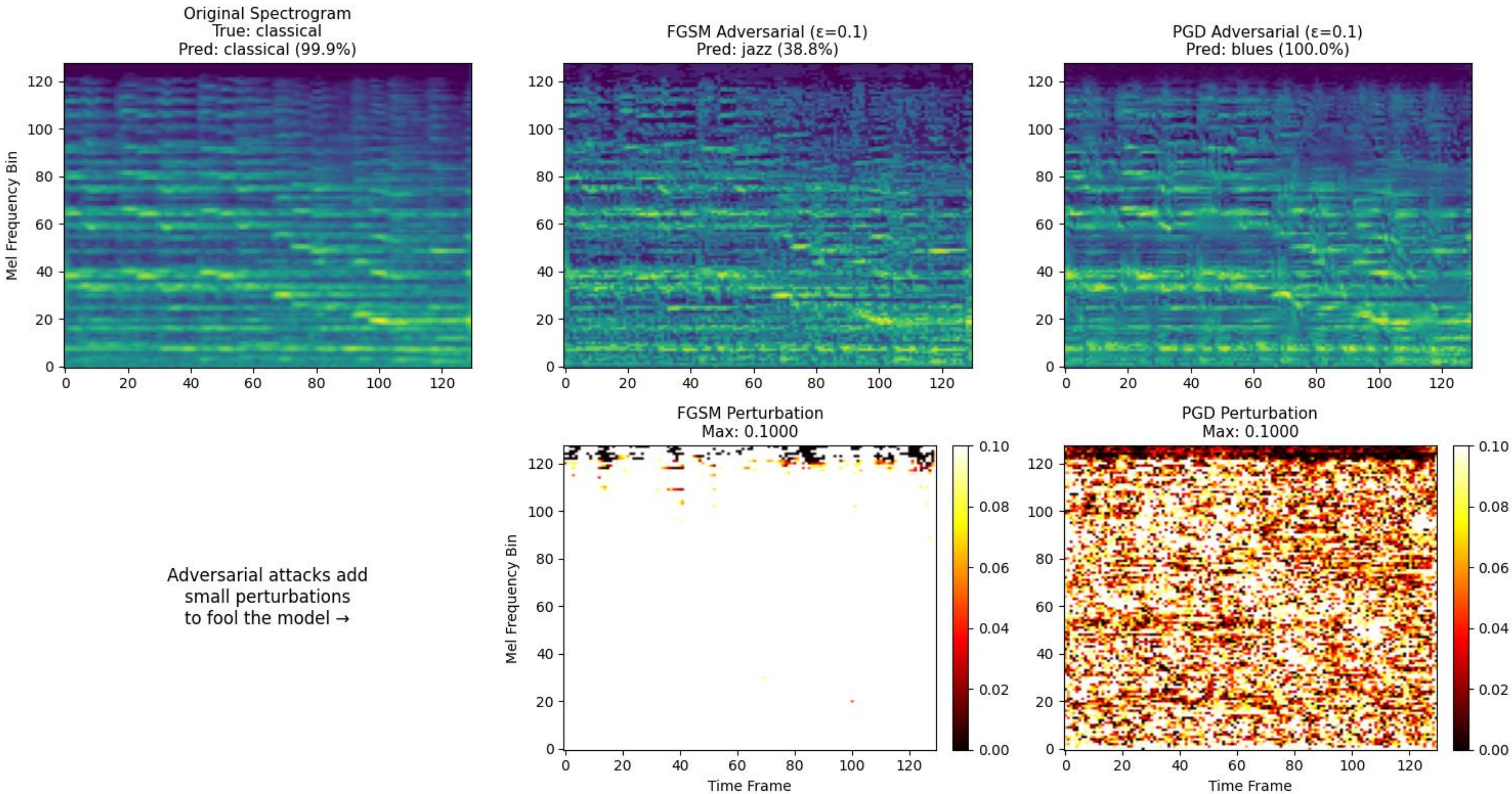
- Single Step Attack
- Moves input in direction of gradient

PGD (Projected Gradient Descent)

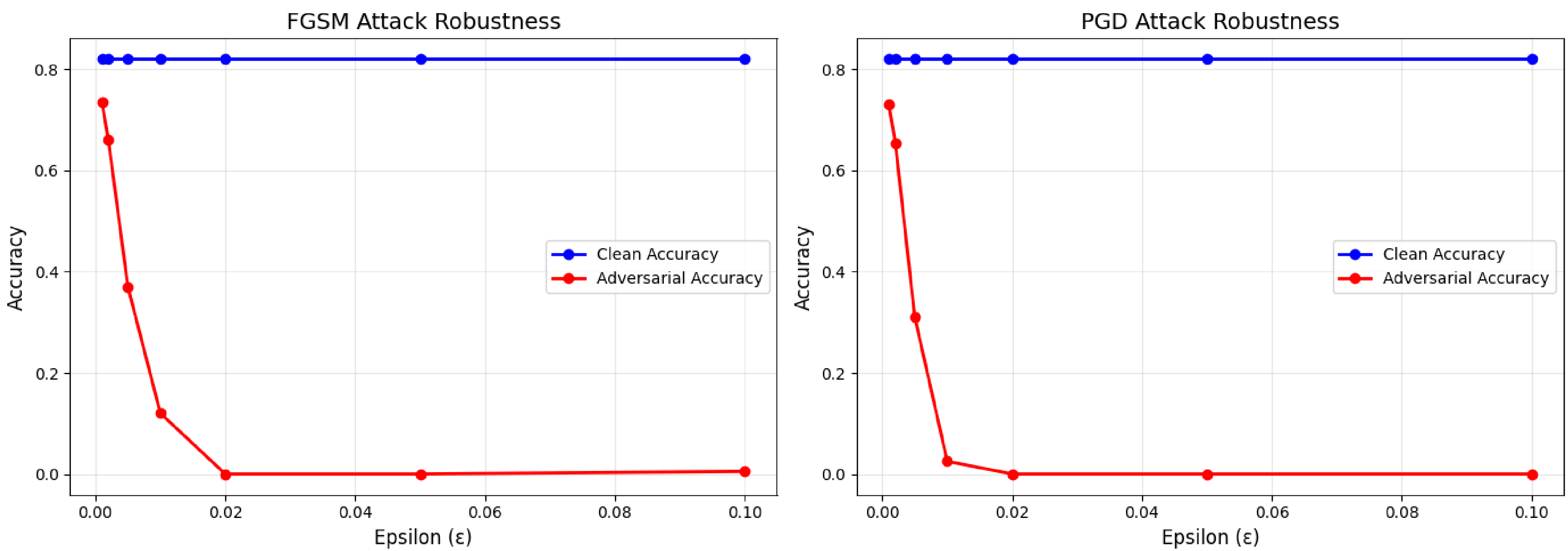
- Multi-step iterative attack
- Takes small steps and projects back into allowed perturbation range

Adversarial Robustness Toolbox

Adversarial Attacks



Adversarial Attacks



GradCAM

1. Baseline Analysis (Correct Predictions)

- 20 correctly classified samples (2 per genre)
- **Goal:** Understand what features the model uses normally

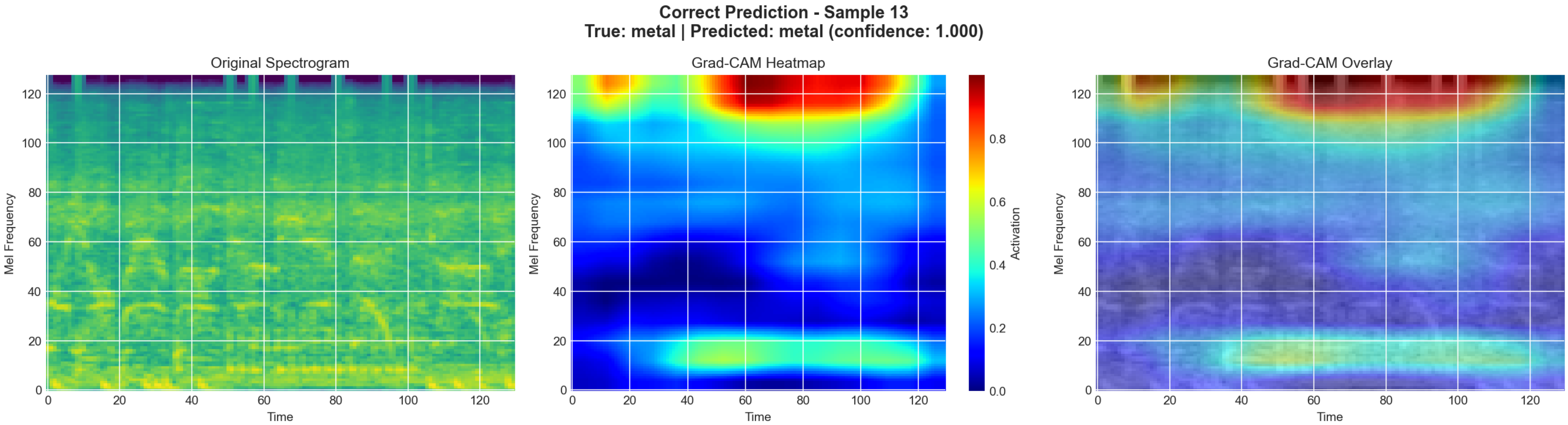
2. Natural Failures (Error Analysis)

- 5 misclassified samples without attacks
- **Goal:** Identify confusion sources

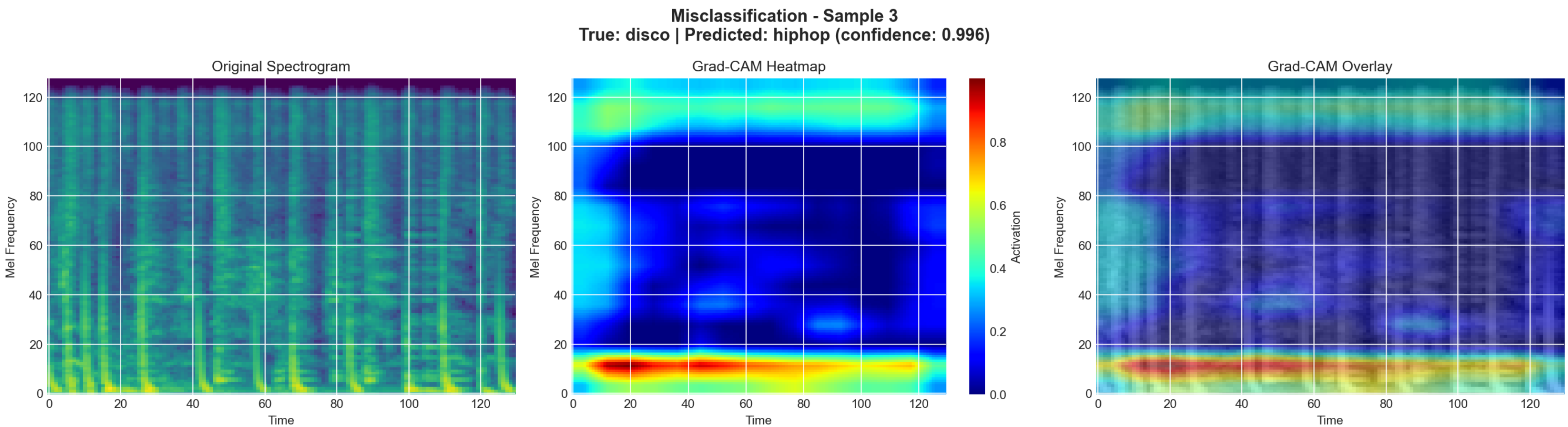
3. Adversarial Analysis (FGSM & PGD, ε=0.1)

- 20 samples under FGSM + 20 under PGD attacks
- Side-by-side comparisons: Clean vs. Adversarial
- **Goal:** How do attacks alter model perception?

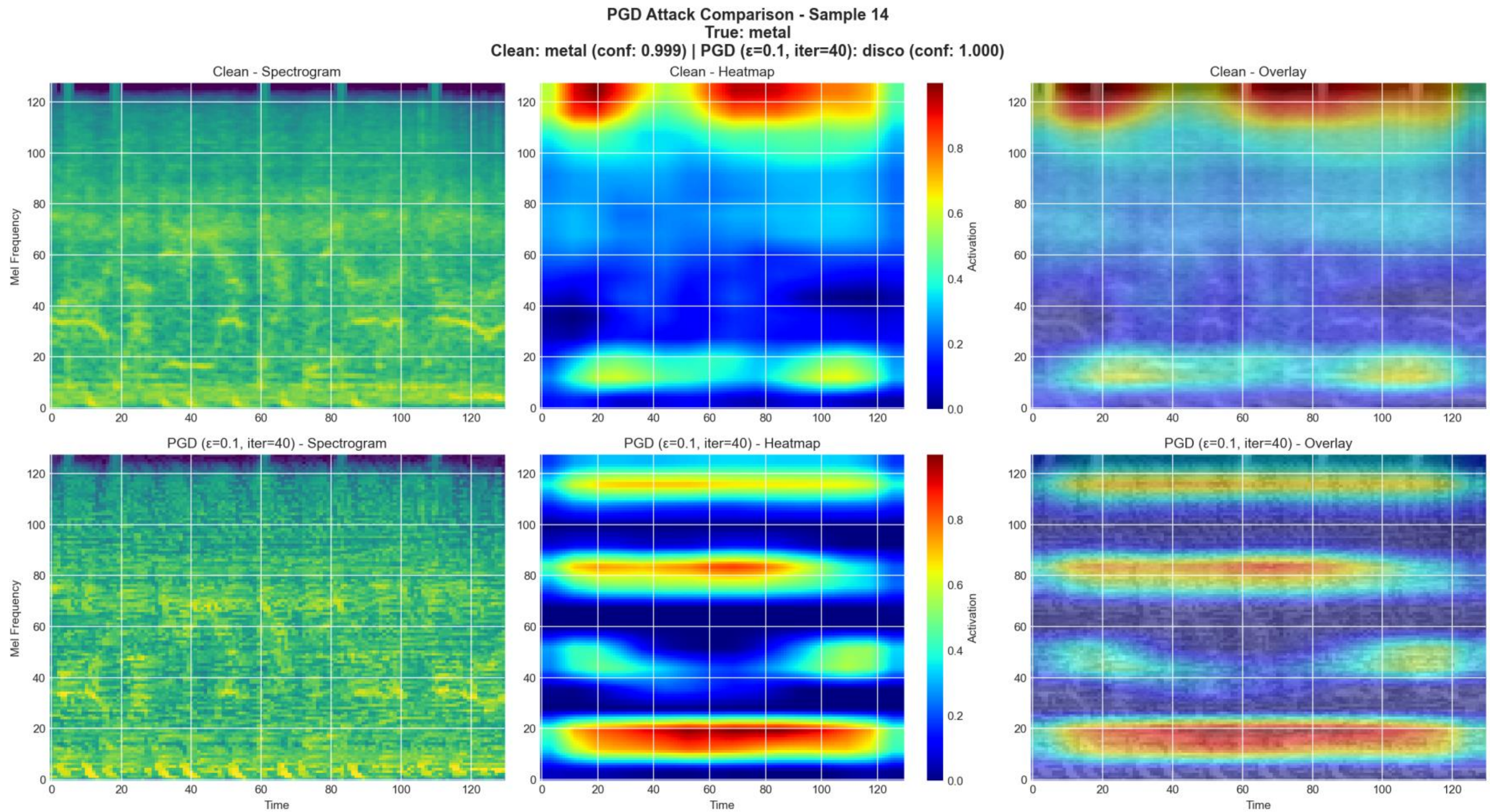
GradCAM – Baseline Analysis



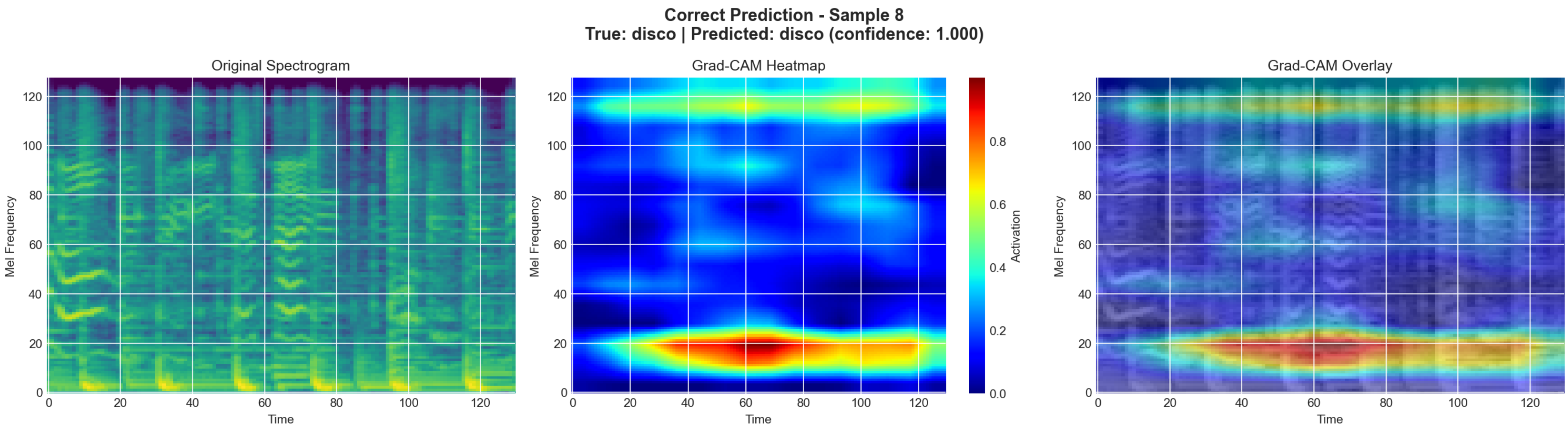
GradCAM – Natural Failures



GradCAM – Adversarial Attacks



GradCAM – Adversarial Attacks



Conclusion

Key Findings	Insight
Model accuracy: 84% song-level	Good baseline performance
Adversarial vulnerability: ~100% at $\varepsilon=0.02$	Highly vulnerable without defense mechanisms
Grad-CAM on clean data: Meaningful attention	Model learns musically relevant features
Grad-CAM under attacks: Artificial patterns	Attacks exploit attention mechanisms