Adversarial Attacks and Defenses

Music Genre Classification

Students:
Bogdan George Carp
Anca-Maria Găină

Coordinators:
Ș. L. Dr. Ing. Ana-Antonia Neacșu
As. Drd. Ing. Vlad Vasilescu

National University of Science and Technology POLITEHNICA Bucharest
Faculty of Electronics, Telecommunications and Information Technology

December 2025

Table of Contents

Introduction

Methodology: Audio Classification

Data Approach

Dataset: GTZAN (1000 songs, 10 genres) converted to **Mel-Spectrograms**.
Concept: Audio classification treated as an image recognition task (128×128 grayscale).

Model 1: Custom CNN (Baseline)

- **Arch:** 4-Block trained from scratch: (Conv2D $\rightarrow$ BN $\rightarrow$ ReLU $\rightarrow$ MaxPool).
- **Complexity:** Filters 32 $\rightarrow$ 256, feeding a 512-unit dense layer.

Model 2: ResNet18 (Transfer)

- **Arch:** Standard ResNet18 (ImageNet).
- **Adaptation:** Modified 1st layer (1-channel) & final layer (10 classes).
- **Role:** SOTA architecture utilizing residual connections.

Training Strategy

- **Optimization:** Adam (LR $= 0.001$) with Weight Decay ($1e^{-4}$).
- **Scheduler:** ReduceLROnPlateau (Factor=0.5, Patience=5).
- **Lifecycle:** Max 100 Epochs. Early Stopping (Patience=20).

Baseline

Model	Clean Accuracy	Training Strategy	Total Training Time
CNN	83.75%	100 Epochs (Early Stop ∼45)	∼11 mins
ResNet18	84.14%	100 Epochs (Early Stop ∼40)	∼13 mins

Tabela 1: Comparison of Model Performance and Training Time

Attacks

Fast Gradient Sign Method (FGSM)

- **Method**: "White Box" attack that uses the model's own gradients.
- **Goal**: To force the music classifier to make a mistake without destroying the audio quality.
- **Calculate Gradient**: Determine which pixels in the Mel Spectrogram contribute most to the correct prediction.
- **Determine Direction**: Find the mathematical direction that increases the error (Loss) the fastest.
- **Add Noise**: Apply a tiny, invisible layer of noise (ϵ) in that exact direction.
- **Result**: The audio sounds the same to the human ear, but the model crosses the decision boundary and misclassifies the genre.

Minimum-Norm Attack

- **Method**: Unlike FGSM (which takes one fixed step), this is an iterative optimization process.
- **Goal**: To fool the genre classifier using the absolute smallest amount of noise possible, making the attack much harder to detect than FGSM.
- **Locate Boundary**: The algorithm analyzes the model to find the closest "decision boundary".
- **Shortest Path**: It calculates the shortest perpendicular vector needed to push the spectrogram just barely over that line.
- **Iterate**: It repeatedly adjusts the input, inching closer to the boundary until the prediction flips with minimal change.
- **Result**: The adversarial noise is mathematically optimized to be as quiet as possible—often completely invisible on a spectrogram and inaudible in the waveform.

Results after attack

Epsilon (ϵ)	CNN (FGSM)	CNN (PGD)	ResNet18 (FGSM)	ResNet18 (PGD)
0.000	81.6%	81.6%	78.4%	78.4%
0.001	79.2%	79.3%	76.5%	76.4%
0.005	66.2%	64.5%	66.2%	66.0%
0.010	52.3%	48.8%	55.3%	53.3%
0.050	17.3%	2.1%	6.9%	0.5%
0.100	9.2%	0.0%	0.5%	0.0%

Tabela 2: Model Robustness under FGSM and PGD Attacks with Varying Epsilon (ϵ)

Defenses

Adversarial Training

- **Method**: Instead of training only on clean GTZAN songs, we train on a mix of clean and attacked data.
- **Goal**: To induce model invariance against gradient-based (FGSM) and optimization-based attacks, effectively hardening the classifier.
- **Dynamic Generation**: Synthesizes adversarial examples on-the-fly via an "inner maximization" step within the training loop.
- **Label Consistency**: Mapping the distorted inputs to their original class labels, compelling the model to maintain correct classification despite significant feature space distortion.
- **Feature Regularization**: Prioritizes structurally invariant features (e.g., rhythm, timbre) over non-robust high-frequency spectral artifacts.
- **Result**: The model learns to recognize the genre even when the attacker tries to confuse it, creating much smoother decision boundaries.

Feature Squeezing

- **Goal**: To detect adversarial inputs by exploiting the fragility of perturbations to signal processing compared to robust natural data.
- **Input Transformation**: Applies non-differentiable transformations to the spectrogram that destroy minute adversarial noise patterns without altering the coarse semantic audio content.
- **Discrepancy Analysis**: Compares prediction vectors between original and "squeezed" inputs to quantify model divergence.
- **Artifact Suppression**: Filters non-robust high-frequency perturbations, forcing adversaries to employ perceptually distinct distortions
- **Result**: Enables the rejection of adversarial samples when the prediction difference exceeds a calibrated threshold, effectively validating input integrity.

Results after defense

Model	Strategy	Training Method	Inference Defense	Clean Acc	Net Diff
CNN	Baseline	Standard	None	81.6%	—
CNN	Adv. Mixed	50% Clean / 50% Adv	None	77.4%	-4.2%
CNN	Adv. Pure	100% Adv	None	75.8%	-5.8%
CNN	Squeezing (Mixed)	50/50 + 5-bit Quant.	5-bit Squeezing	70.9%	-10.7%
CNN	Squeezing (Pure)	100% Adv + 5-bit Quant.	5-bit Squeezing	70.8%	-10.8%
ResNet18	Baseline	Standard	None	78.4%	—
ResNet18	Adv. Mixed	50% Clean / 50% Adv	None	73.3%	-5.1%
ResNet18	Adv. Pure	100% Adv	None	74.1%	-4.3%
ResNet18	Squeezing (Mixed)	50/50 + 5-bit Quant.	5-bit Squeezing	74.1%	-4.3%
ResNet18	Squeezing (Pure)	100% Adv + 5-bit Quant.	5-bit Squeezing	—	—

Tabela 3: Impact of Defense Strategies on Clean Accuracy and Performance Drop

Results

Detailed Robustness Analysis (FGSM vs PGD)

	Clean	$\epsilon = 0.01$		$\epsilon = 0.03$		$\epsilon = 0.1$	
Strategy	**Acc**	**FGSM**	**PGD**	**FGSM**	**PGD**	**FGSM**	**PGD**
Model: CNN							
Baseline	81.6%	52.3%	48.8%	**26.2%**	**8.5%**	**9.2%**	0.0%
Adv. Mixed	77.4%	**68.4%**	**67.8%**	52.9%	48.6%	**11.7%**	**2.3%**
Adv. Pure	75.8%	**67.2%**	**67.0%**	51.0%	47.7%	**15.8%**	**6.4%**
Squeezing (Mix)	70.9%	60.9%	60.8%	46.6%	44.3%	**14.5%**	**4.8%**
Squeezing (Pure)	70.8%	62.7%	62.6%	48.3%	45.3%	**14.9%**	**6.2%**
Model: ResNet18							
Baseline	78.4%	55.3%	53.1%	**21.9%**	**9.3%**	0.5%	0.0%
Adv. Mixed	73.3%	**64.2%**	**64.2%**	44.1%	40.7%	**7.3%**	0.5%
Adv. Pure	74.1%	**66.2%**	**66.2%**	49.1%	47.3%	**10.3%**	**2.9%**
Squeezing (Mix)	74.1%	**64.1%**	**63.8%**	42.4%	39.4%	**9.1%**	**1.4%**
Squeezing (Pure)	74.1%	**64.1%**	**63.1%**	45.6%	42.7%	**10.2%**	**3.0%**

Tabela 4: Comparison of accuracy under FGSM and PGD attacks at varying perturbation strengths (ϵ).

Model	Baseline (Undefended)	Defended (Adv. Mixed)	Defended (Adv. Pure)	Improvement (Mixed)
CNN	0.55	20.32	34.43	37x
ResNet18	0.55	4.85	6.36	9x

Tabela 5: Defense Improvement Summary

Best performing Defence

Model	Best Defense Strategy	Epsilon (ϵ)	Baseline Acc	Defended Acc	Improvement
CNN	Adv. Training (Mixed)	0.01	48.8%	67.8%	+19.0%
CNN	Adv. Training (Mixed)	0.03	8.5%	48.6%	+40.1%
ResNet18	Adv. Training (Pure)	0.01	53.1%	66.2%	+13.1%
ResNet18	Adv. Training (Pure)	0.03	9.3%	47.3%	+38.0%

Tabela 6: Summary of Improvements with Best Defense Strategies

Conclusions

Conclusions

Bibliography I

Corey Kereliuk, Bob L. Sturm, and Jan Larsen.
Deep learning and music adversaries.
IEEE Transactions on Multimedia, 17(11):2059–2071, 2015.

Yijie Xu and Wuneng Zhou.
A deep music genres classification model based on cnn with squeeze excitation block.
In *Proceedings, APSIPA Annual Summit and Conference 2020*, 2020.

Yunming Liang, Yi Zhou, Tongtang Wan, and Xiaofeng Shu.
Deep neural networks with depthwise separable convolution for music genre classification.
In *2019 2nd IEEE International Conference on Information Communication and Signal Processing*, 2019.

Nitin Choudhury, Deepjyoti Deka, Parismita Sarma, and Satyajit Sarmah.
Music genre classification using convolutional neural network.
In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, 2023.

Bibliography II

Nikki Pelchat and Craig M. Gelowitz.
Neural network music genre classification.
In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, 2019.