Adversarial Attacks Applied to Whale-Detecting Neural Network

Jerry Kurtin – Georgetown High School – Georgetown, TX
Supervisors: Reid Wyde and Scott Johnston
Signal and Information Sciences Laboratory

Background

- Modern neural networks tend to be susceptible to adversarial attacks.
- Adversarial attack: a small, targeted disruption to an input image that causes a model to misclassify the image
- Adversarial attacks could cause real-world damage as important technology begins to rely on machine learning.

Objective

Dataset: 30,000 2-second audio clips from ocean buoys run by Cornell University

- Create a neural network that can distinguish North Atlantic right whale calls from ocean noise and other whale calls
- Discover vulnerabilities in the model through white-box and black-box attacks

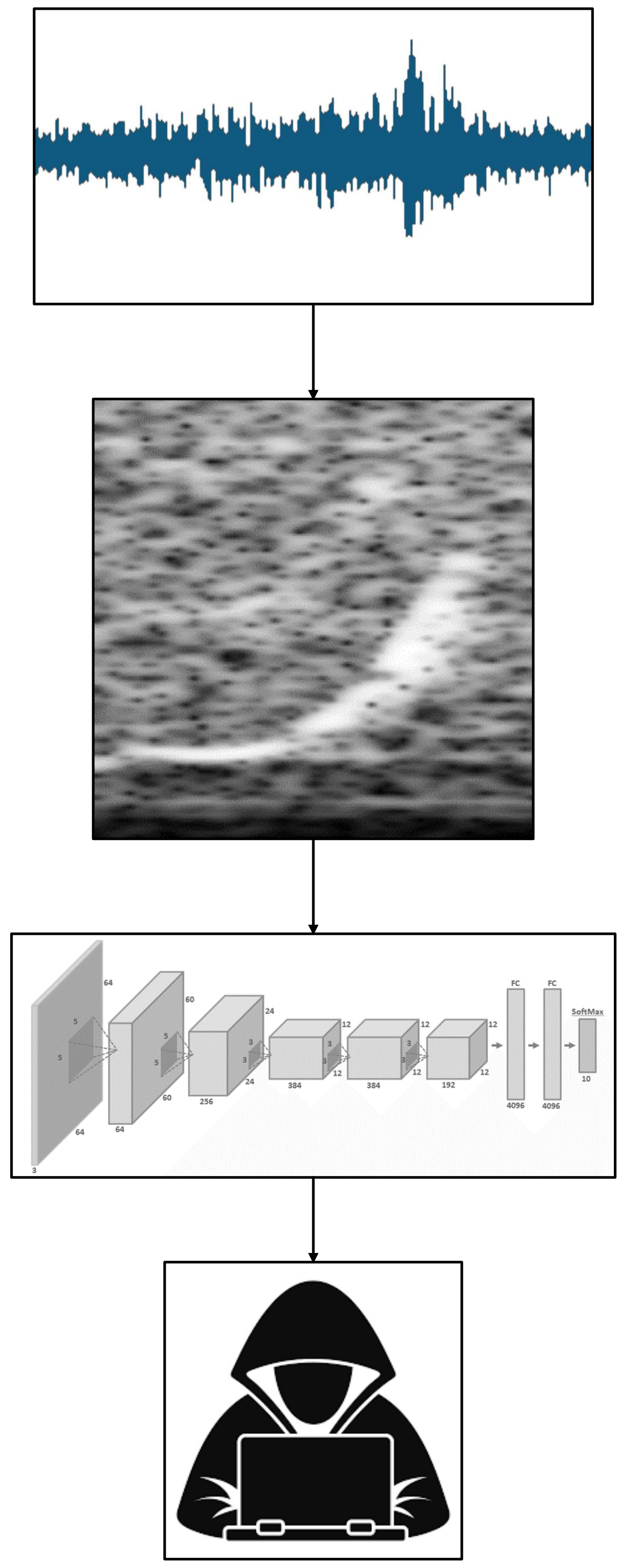
Process

Convert audio clips to spectrograms

Train AlexNet, a convolutional neural network (CNN), on audio clips

Alter images using adversarial attacks

Test new images for reduced performance

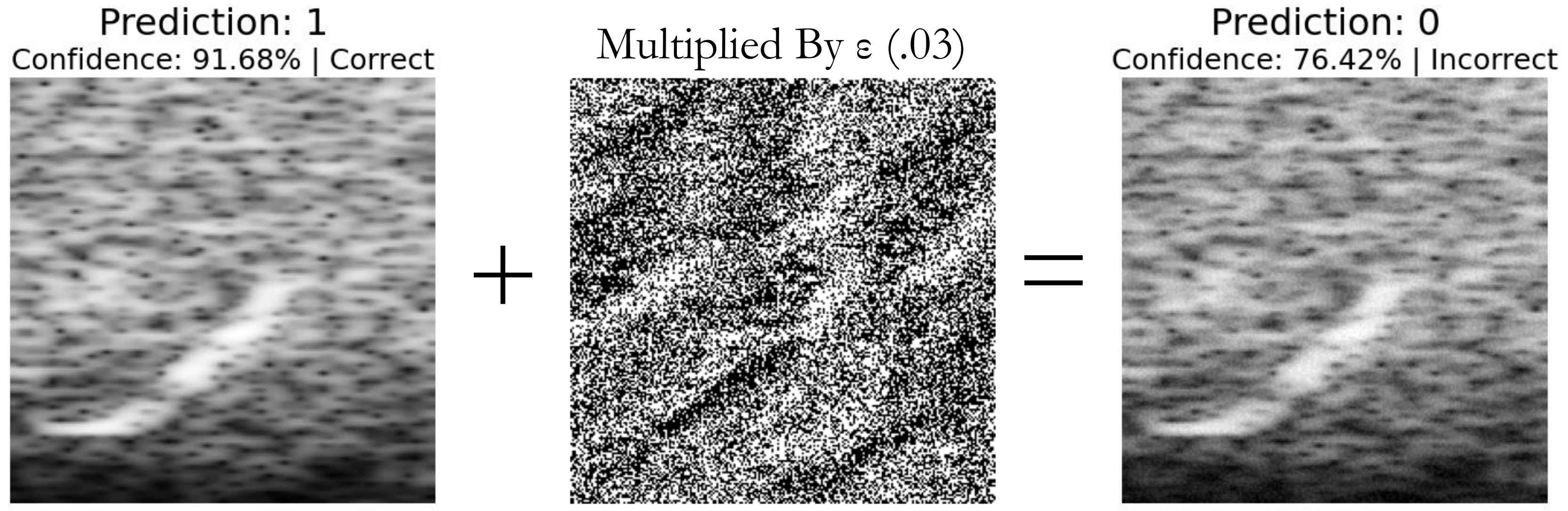


Attack | Original Image | Perturbation | Perturbed Image

White Box Attacks: unrestricted access to model

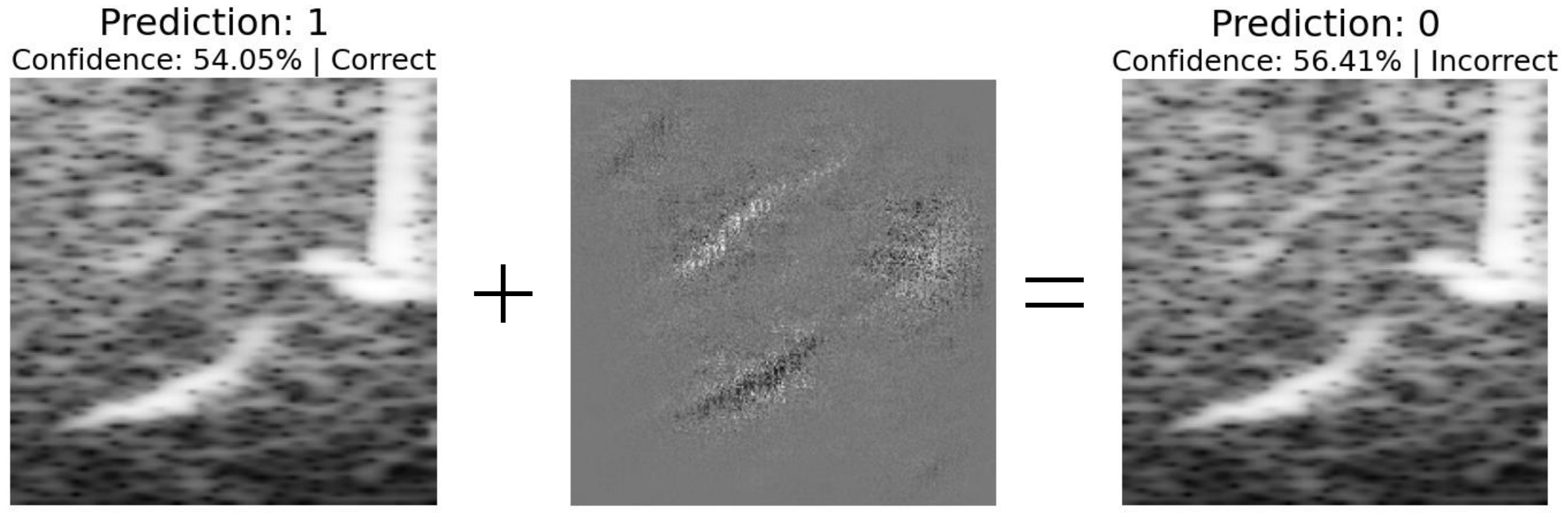
Fast Gradient Sign Method (FGSM)

- Calculates model gradient on image
- Steps in the opposite direction by a constant ε
- Computationally cheap, but noticeable



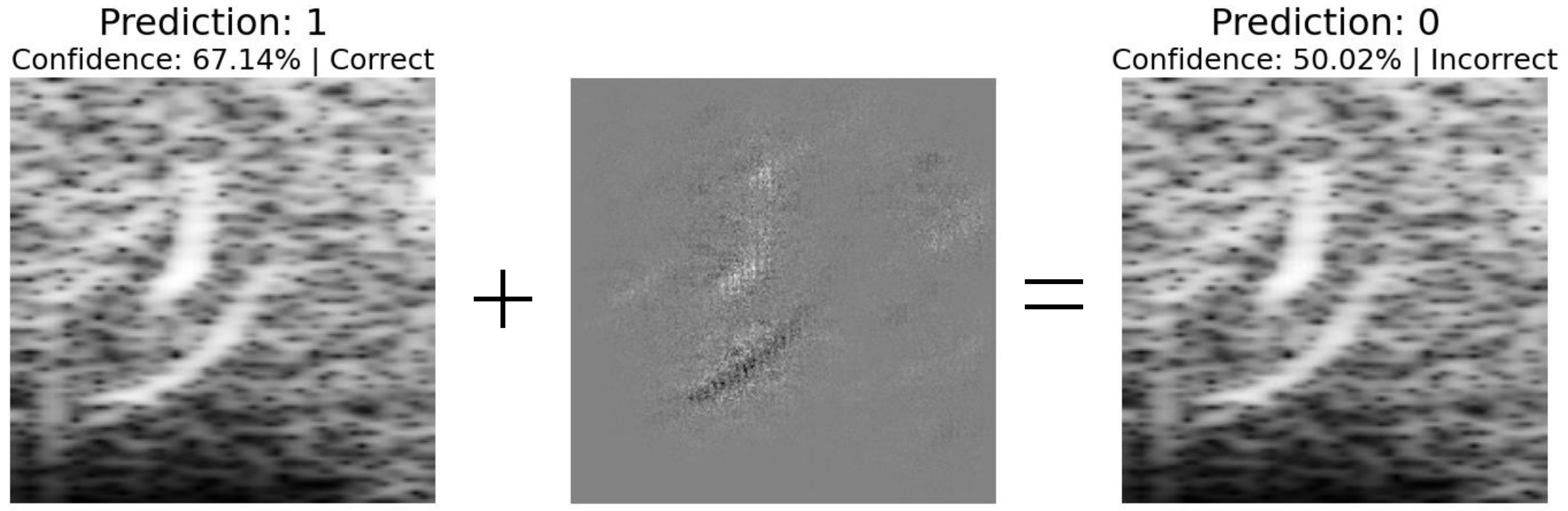
Carlini and Wagner (CW)

- Searches for smallest change to image that causes misclassification
- 1000 steps taken
- Computationally expensive, but more effective and less noticeable



DeepFool

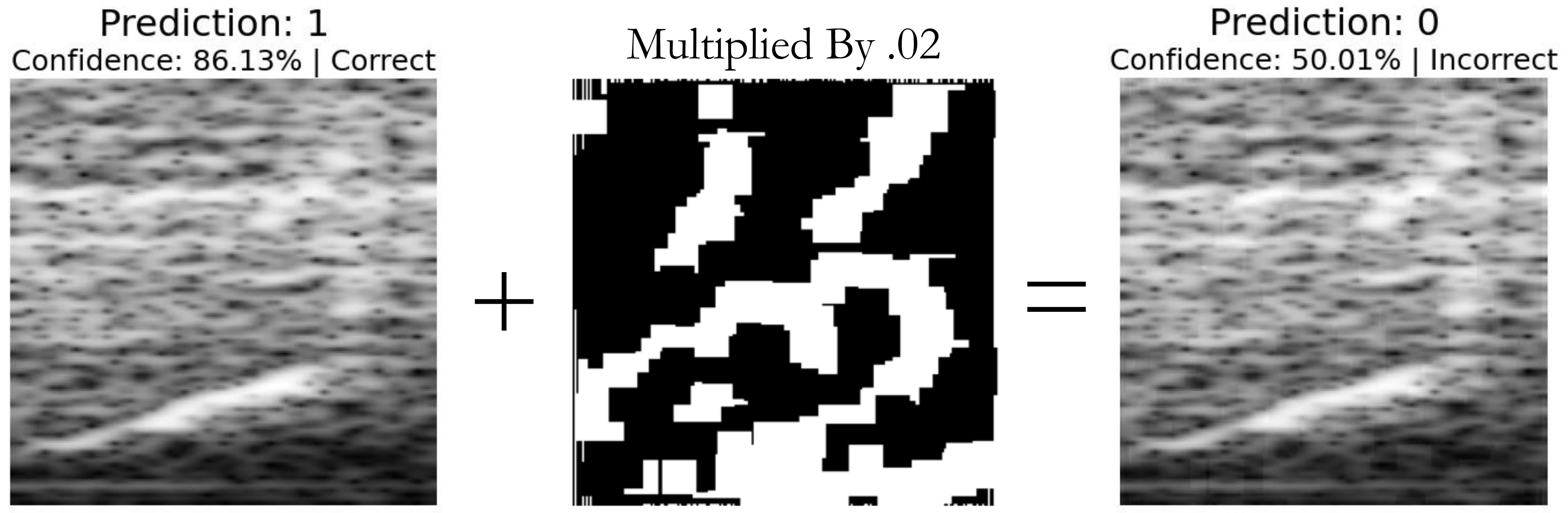
- Finds nearest hyperplane
- Calculates the changes needed to cross the hyperplane
- Hyperplane: a high-dimensional 'line' separating different classifications
- Efficient and subtle



Black Box Attacks: Only given access to a model's final decision and certainty

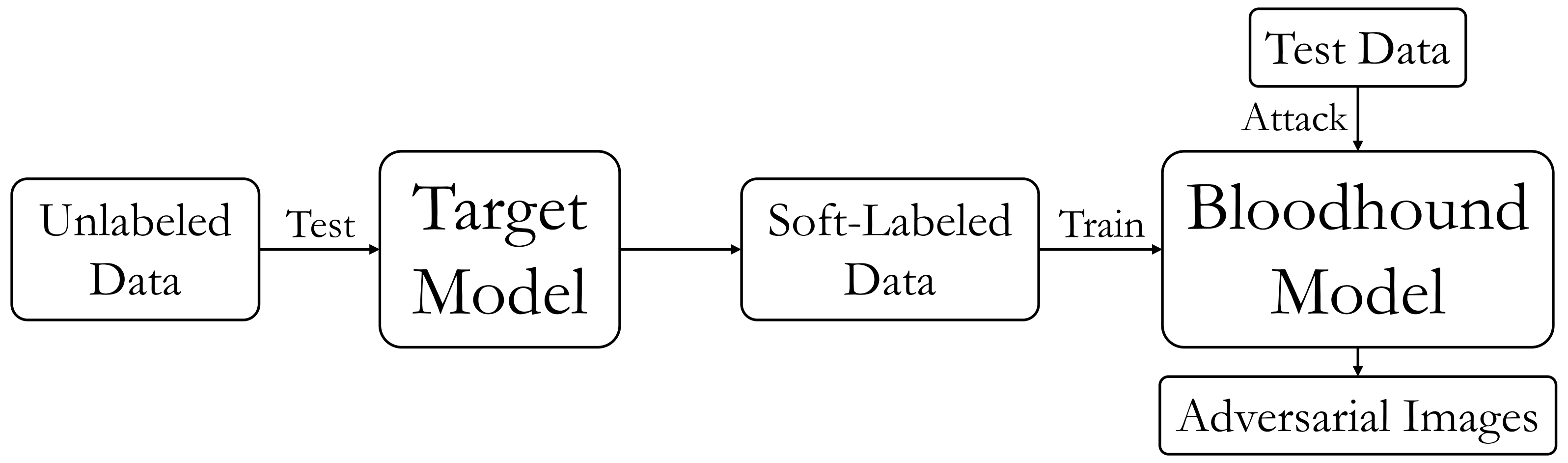
Square

- Changes a random square of pixels
- Tests for reduced certainty in model
- Repeats until successful misclassification
- "Guess and check"

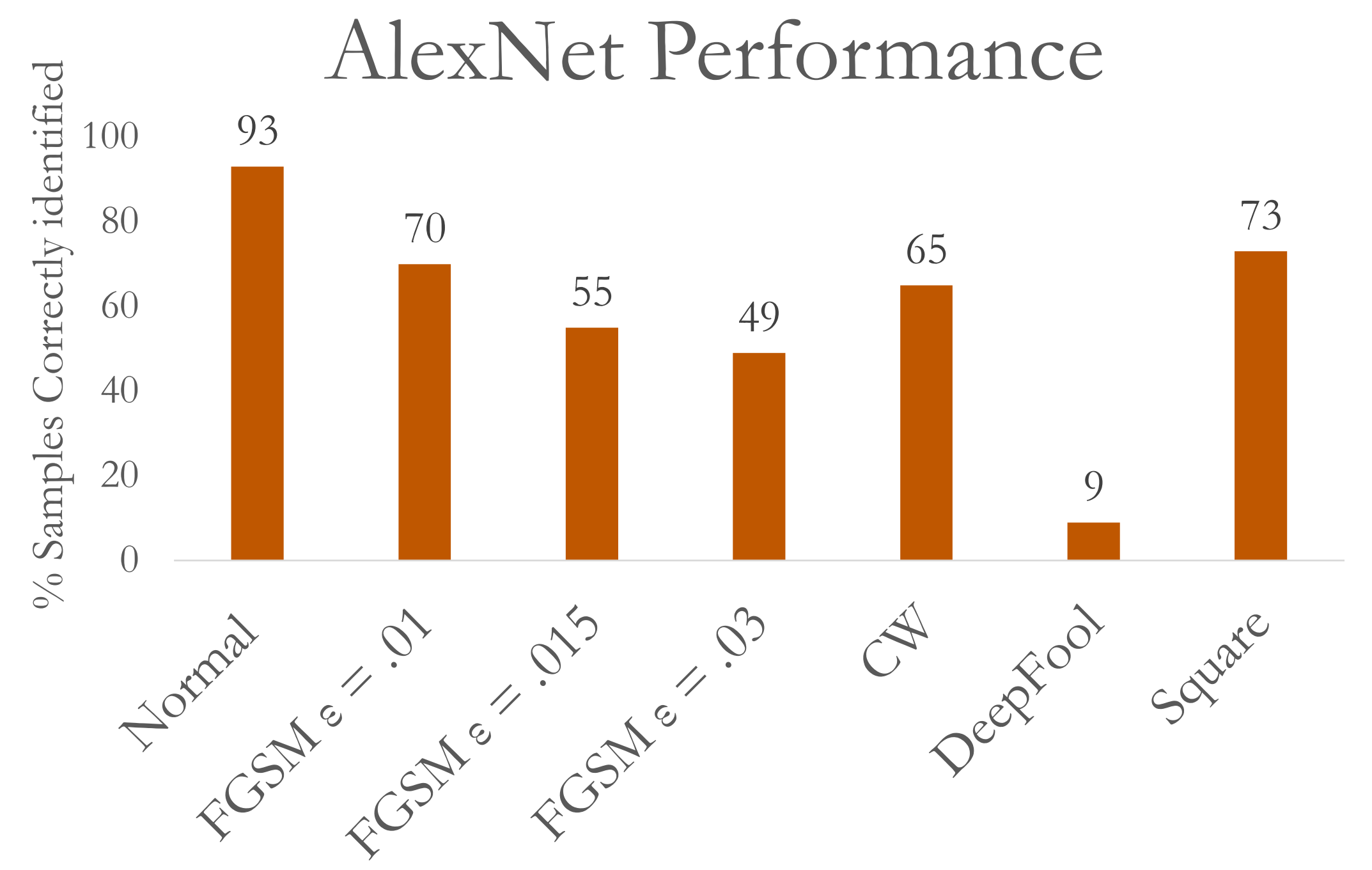


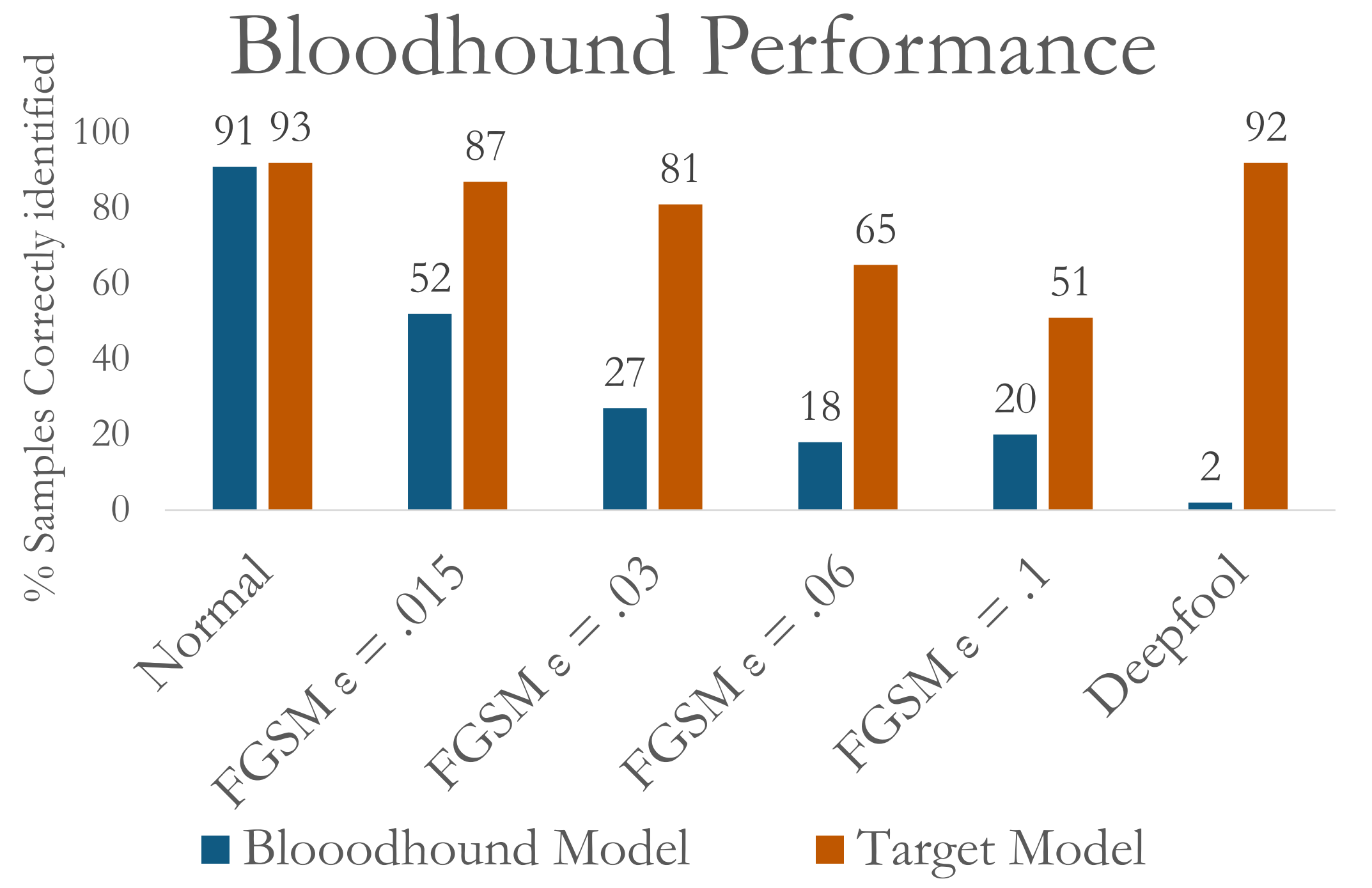
Bloodhound

- Labels spectrograms with output of target model
- Trains a 'bloodhound' model on labeled outputs
- Performs white-box attacks on bloodhound model


Unlabeled Data
Test
Target Model
Soft-Labeled Data
Train
Bloodhound Model
Test Data
Attack
Adversarial Images

Results





Conclusion

- Image-recognition CNNs can be accurately used for sound classification
- White and black box attacks succeeded in reducing accuracy below random chance
- Decision borders are cloudy due to small dataset

Moving Forward

- Bootstrap dataset to train generalization
- Create realistic attacks that perturb original sound samples
- Expand network to detect and identify animal calls and human activity

Acknowledgements

I'd like to thank Reid Wyde, Scott Johnston, Anna Chaney, and Hector Gonzalez for their generous mentorship and patience.