Music to My Ears: Turning GPU Sounds into Intellectual Property Gold

Sayed Erfan Arefin[1], Abdul Serwadda[1]

[1]Texas Tech University, Lubbock, Texas, United States.

Introduction

Deep Neural Networks (DNNs) are critical to modern applications, making their architecture a valuable asset. Side-channel attacks exploit indirect data leaks to uncover sensitive information. We introduce a novel acoustic side-channel attack that captures GPU sound emissions during DNN operations.

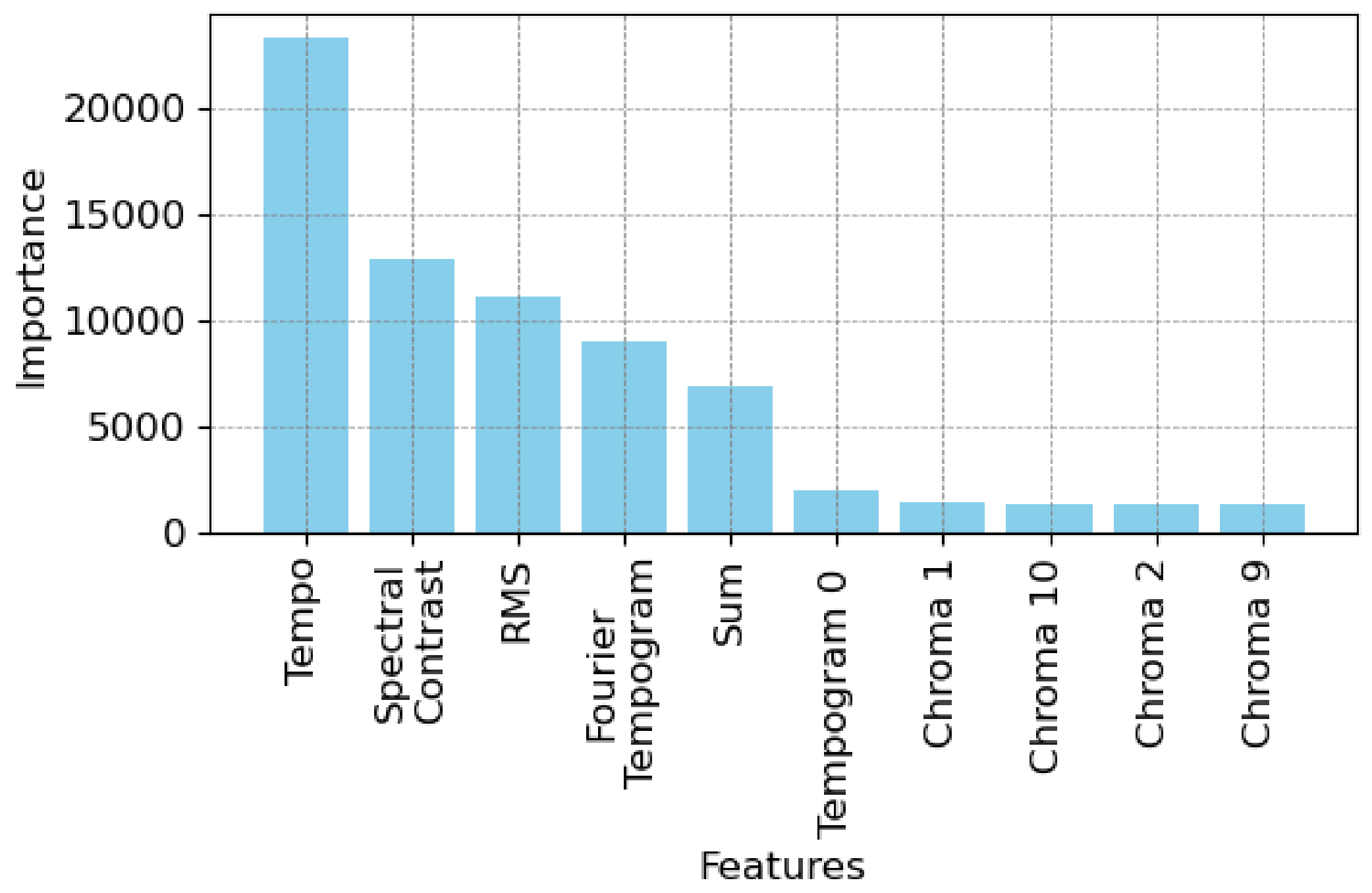
Using a MEMS microphone, we show that distinct sound patterns from GPU components can be analyzed to infer the architecture of DNNs. Our experiments on prominent models like ResNet and VGG demonstrate the practicality of this attack, revealing a significant vulnerability in DNN security.

Experiment Setup

1. A total of 24 Convolutional Neural networks were involved in experimentation. These were categorized into 10 core models and their variants.
2. We provided inputs to these CNNs and collected acoustic data. There are 4 types of input datasets in our experiment. These are:
   (a) A single grayscale image, tested 100 times across all configurations.
   (b) The same image as above, but in color, tested 100 times.
   (c) 50 diverse images of cats and dogs, widely used as a neural network benchmark.
   (d) 100 randomly selected colorful images, tested across all configurations.
3. We experimented with 3 different configurations. These are:
   (a) **Configuration 1:** The attacker has access to the training data for 10 core models and tries to classify the victimâs core architecture.
   (b) **Configuration 2:** The victim uses a variant of a core architecture. The attacker attempts to infer the architectural family of the variant using data collected from the core models.
   (c) **Configuration 3:** The attacker has access to the variants of the core architectures and attempts to identify the specific variant used by the victim.
4. Data collection setup for the desktop environment is presented below.

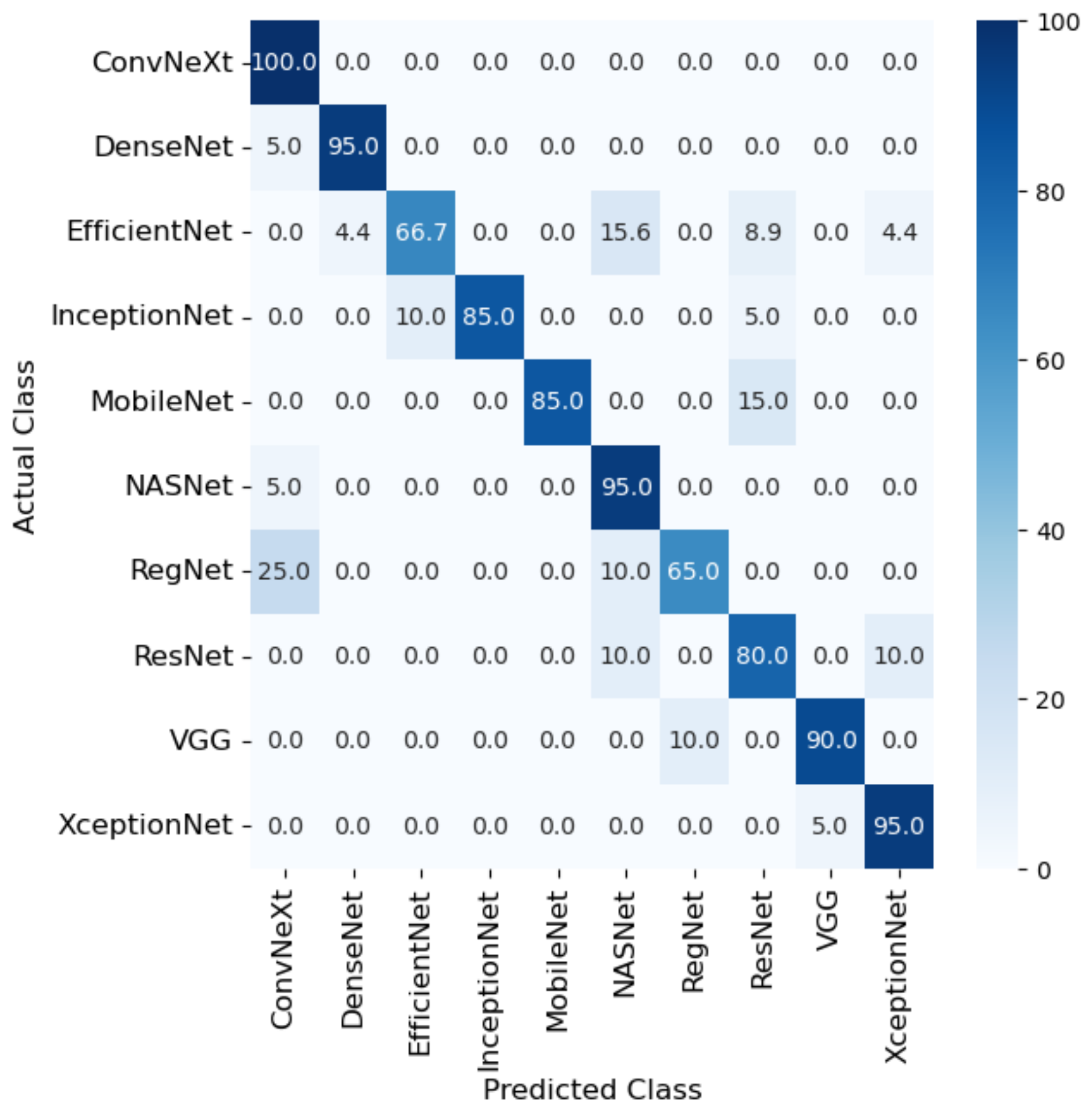
Features Analysis

- **Statistical Features:** These include measures like mean, variance, and standard deviation of the acoustic signals. Calculated using NumPy.
- **Spectral Features:** These were calculated using the Librosa library. Some key features are: Spectral Centroid, Fourier tempogram, Spectral Contrast etc.
- The plot below shows feature importance analysis of one of the configurations: Configuration-1 for Dataset-1 on Nvidia Jetson Nano



Results

1. We ran classification on the extracted features for each experiment configurations. Classifiers included a custom neural network, K-Nearest Neighbour, Random Forest.
2. The Confusion matrix for the configuration-1 for Dataset-4 on RTX 2060 in presented below.

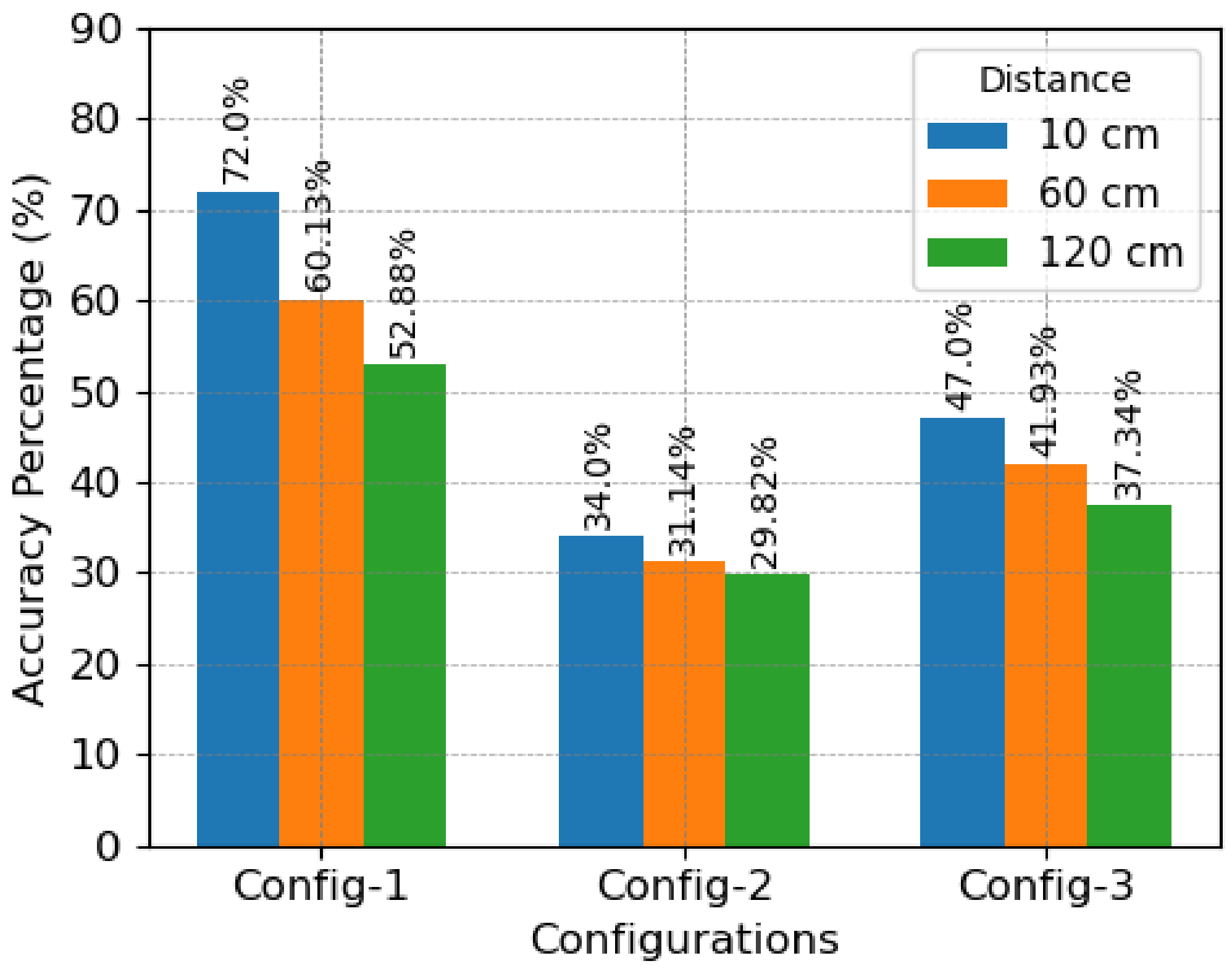


3. The performance of our acoustic side-channel attack was evaluated across different configurations, devices, and datasets. The key findings of our acoustic side-channel attack are summarized below:
   - **Configuration 1: Core Model Identification**
     - **Jetson Nano:** 95% accuracy (Dataset-1)
     - **RTX 2060:** 85.50% accuracy (Dataset-4)
   - **Configuration 2: Architectural Family Identification**
     - **RTX 2060:** 42.14% accuracy (Dataset-2)
     - **Jetson Nano:** 37% accuracy (Dataset-1)
   - **Configuration 3: Variant Identification**
     - **Jetson Nano:** 85.29% accuracy (Dataset-1)
     - **RTX 2060:** 68.75% accuracy (Dataset-4, Random Forest classifier)

Sensitivity Analysis

We tested the impact of microphone distance from the GPU to understand how it affects the attackâs accuracy. The microphone was placed at 10 cm, 60 cm, and 120 cm distances.

- As distance increased, the attackâs accuracy dropped from 85.50% at 10 cm to 49.00% at 120 cm.
- Greater distances reduced the SNR, making it harder to extract meaningful acoustic data. However, the attack remains effective even at 120 cm.
- Interestingly, the accuracy drop between 10 cm and 60 cm is more significant than the drop between 60 cm and 120 cm. This can be attributed to interference patterns that emerge when sound waves reflect off nearby surfaces, creating areas of constructive and destructive interference. These phenomena can amplify or diminish the recorded acoustic signals at certain distances, influencing the attackâs effectiveness.
- The graph below shows the neural networkâs accuracy at various microphone distances from the target machine, averaged across the dataset.



Conclusions

Our study presents the following key findings:

- **Novel Acoustic Side-Channel Attack:** We demonstrated a new attack method that leverages GPU acoustic emissions to infer DNN architectures, revealing a significant security vulnerability.
- **Feasibility and Validation:** Using a MEMS microphone, the attack was validated on multiple CNN architectures, including ImageNet models, across various GPU devices and various input datasets.
- **Sensitivity Analysis:** The attack adapts well to different conditions, showing strong results in identifying CNN architectures when the training data contains similar models.
- **Impact of System Processes:** While system processes were left running during experiments, the controlled environment ensured minimal interference. However, additional processes may affect the accuracy of the attack.
- **Needs to Improved Security:** This work highlights the need for enhanced security measures against acoustic side-channel attacks and encourages further research in this area.

References

[1] Weizhe Hua, Zhiru Zhang, and G. Edward Suh. Reverse engineering convolutional neural networks through side-channel information leaks. DAC '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357005. doi: 10.1145/3195970.3196105.

[2] Peng Cheng, Ibrahim Ethem Bagci, Utz Roedig, and Jeff Yan. Sonarsnoop: active acoustic side-channel attacks. *International Journal of Information Security*, 19(2):213–228, Apr 2020. ISSN 1615-5270. doi: 10.1007/s10207-019-00449-8.

Contact information:
Sayed Erfan Arefin
erfanjordison@gmail.com