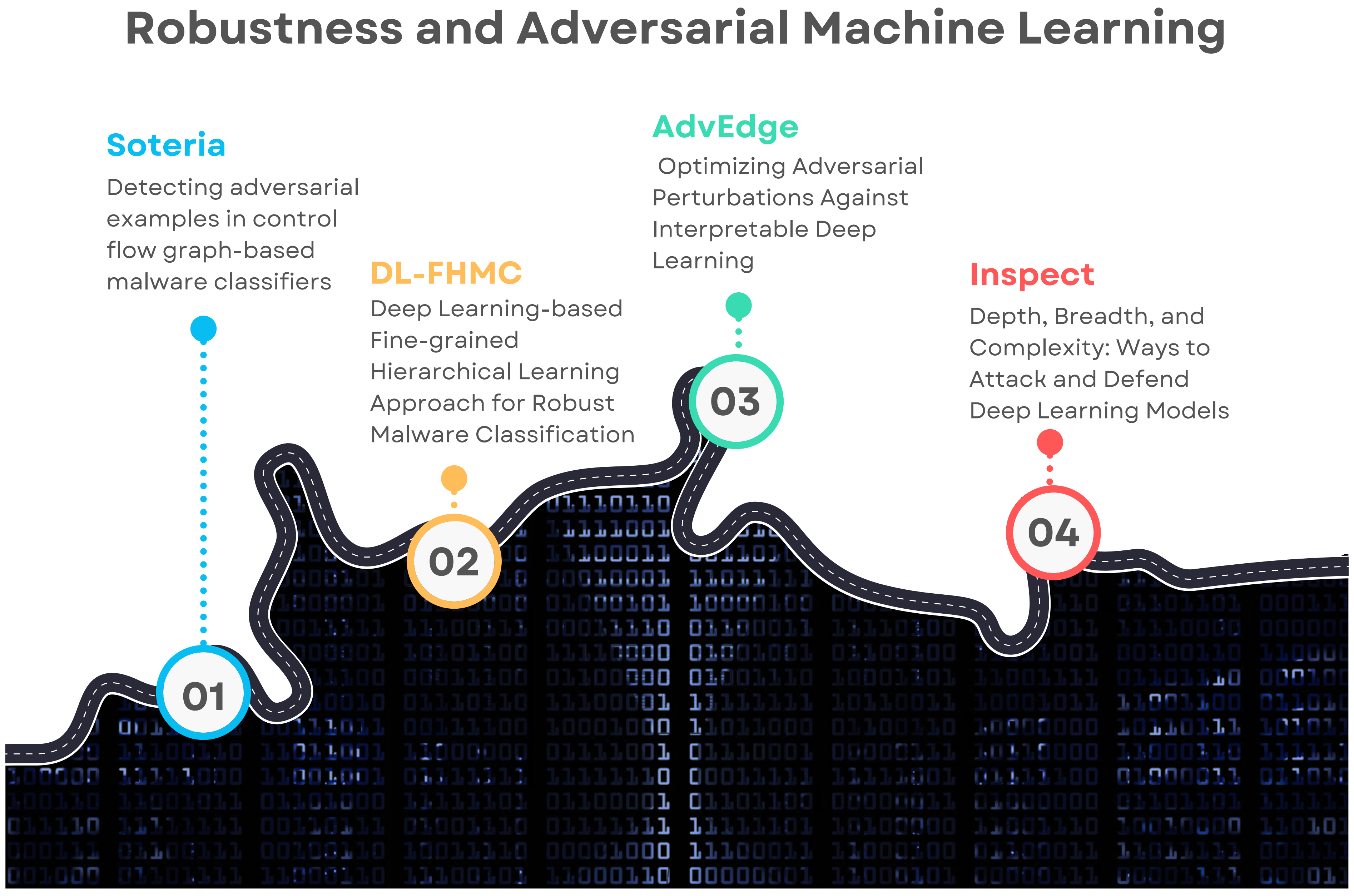
Robustness and Adversarial Machine Learning

Mohammed Abuhamad

Loyola University Chicago

Research Items: Investigating the Security Proprieties of Machine Learning Models



Robustness and Adversarial Machine Learning

01 Soteria
Detecting adversarial examples in control flow graph-based malware classifiers

02 DL-FHMC
Deep Learning-based Fine-grained Hierarchical Learning Approach for Robust Malware Classification

03 AdvEdge
Optimizing Adversarial Perturbations Against Interpretable Deep Learning

04 Inspect
Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models

Soteria

Soteria: Detecting Adversarial Examples in Control Flow Graph-based Malware Classifiers. [3]

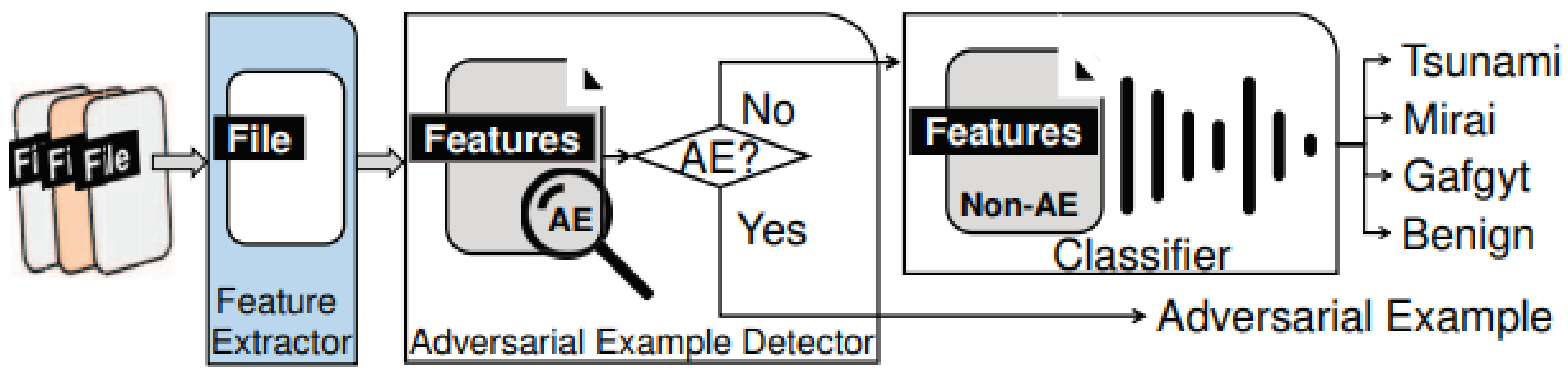


Figure 1. The architecture of Soteria. IoT samples are fed to the feature extraction process, where each sample is represented by multiple feature vectors. The feature vectors are forwarded to adversarial example detector. All non-AEs are then forwarded to the classifier to be classified into its corresponding family

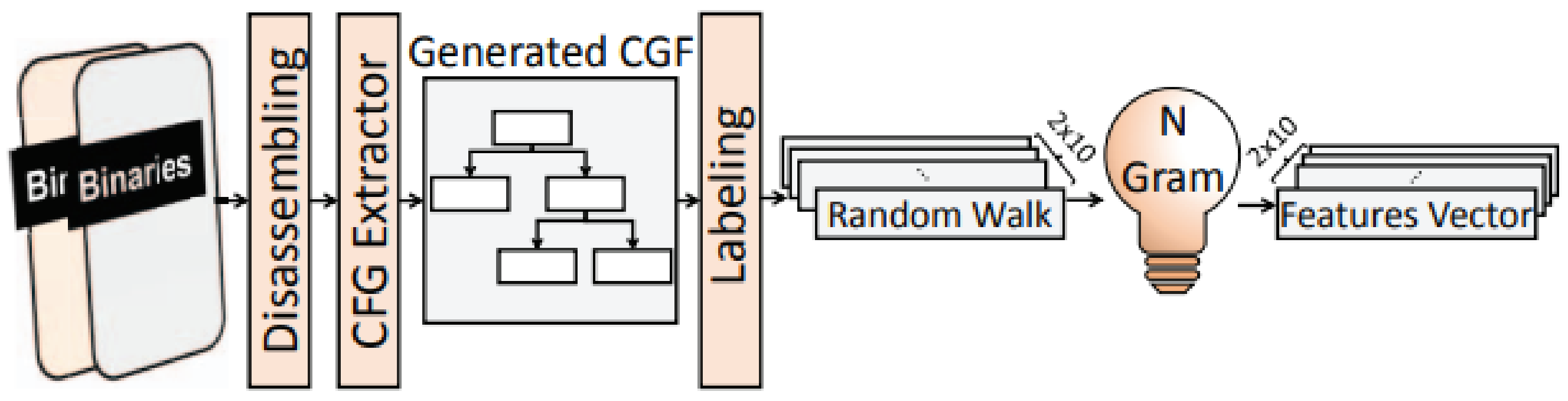


Figure 2. Soteria feature extraction process. IoT samples binaries are disassembled to extract their corresponding CFGs. Then, two nodes labeling techniques are used (Dense-based and level-based), then, several random walks are done over each labeled graph. The trace of the random walk is then used for feature extraction.

DL-FHMC

DL-FHMC: Deep Learning-based Fine-grained Hierarchical Learning Approach for Robust Malware Classification [2]

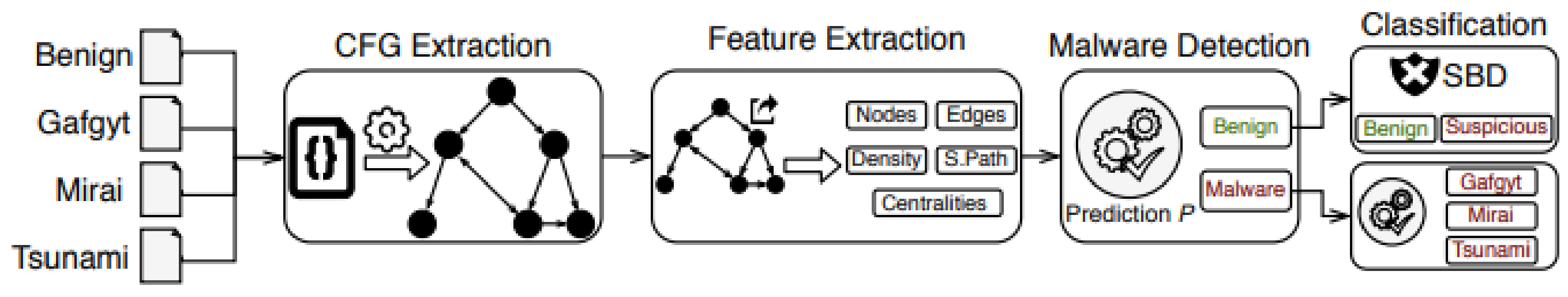


Figure 3. DL-FHMC system flow. First, corresponding CFGs of the IoT software are extracted, then, 23 algorithmic features are extracted from the CFGs. Afterward, an IoT malware detection system classifies samples into benign and malware, all malware samples are directed to IoT malware classification system, while benign samples are directed into suspicious behavior detection system (SBD) for further investigation.

Suspicious behavior detection. The design consists of four modules, a subgraphs mining module to extract frequent subgraphs from three IoT malicious families. The subgraphs are ranked by the pattern selection module. The CFG of each sample is redirected to the Suspicious Behavior Detector as a 30,000-d vector.

AdvEdge

AdvEdge: Optimizing Adversarial Perturbations Against Interpretable Deep Learning [1]. Other exploration can be found in [4].

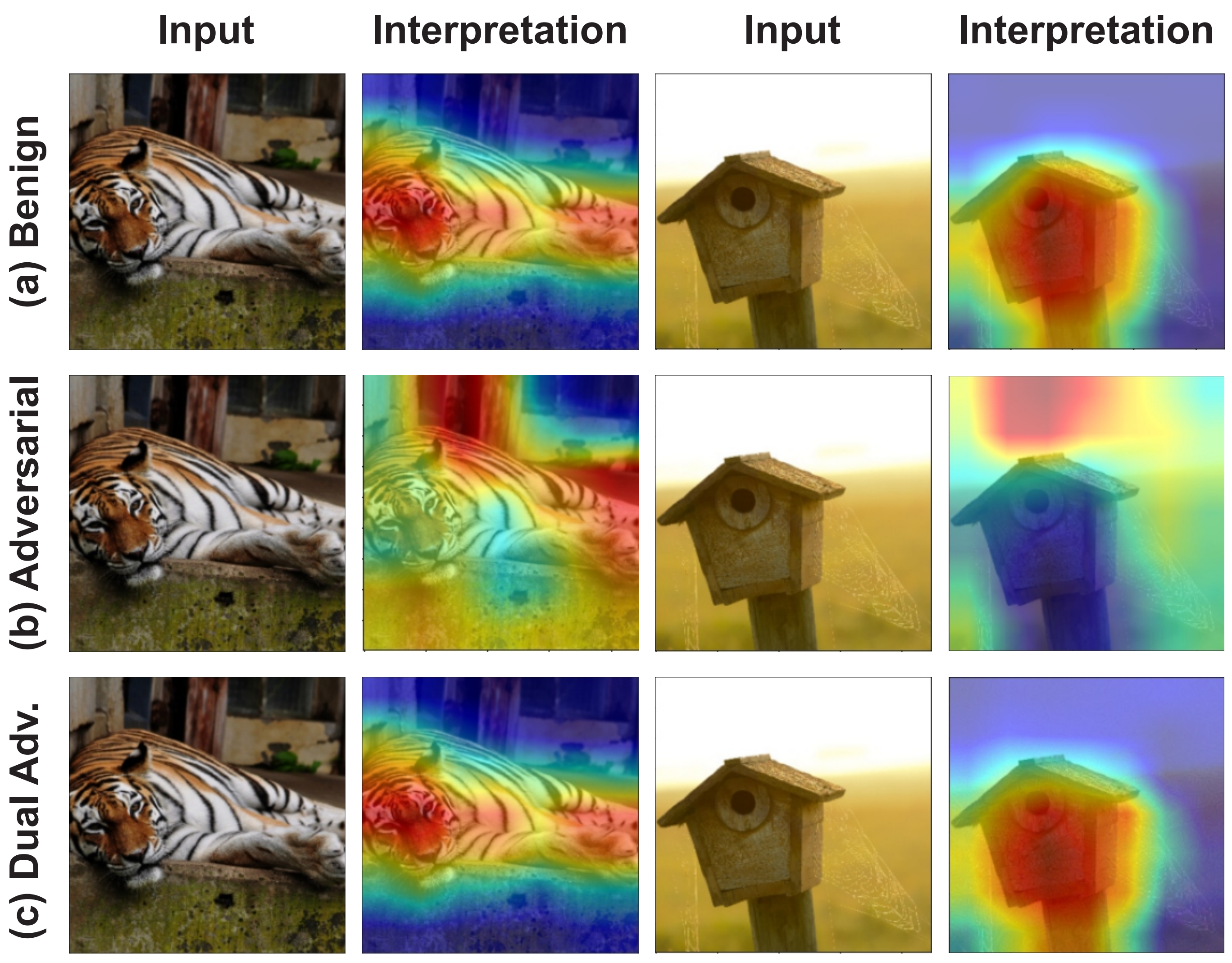


Figure 4. Example images for (a) benign, (b) regular adversarial and (c) dual adversarial and interpretations on ResNet (classifier) and CAM (interpreter).

We present AdvEdge and AdvEdge+, two attacks to mislead the target DNNs and deceive their combined interpretation models. We evaluate the proposed attacks against two DNN model architectures coupled with four representatives of different categories of interpretation models. The experimental results demonstrate our attacks' effectiveness in deceiving the DNN models and their interpreters.

References

[1] Eldor Abdukhamidov, Mohammed Abuhamad, Firuz Juraev, Eric Chan-Tin, and Tamer AbuHmed. Advedge: Optimizing adversarial perturbations against interpretable deep learning. In *International Conference on Computational Data and Social Networks*, pages 93–105. Springer, Cham, 2021.

[2] Ahmed Abusnaina, Mohammed Abuhamad, Hisham Alasmary, Afsah Anwar, Rhongho Jang, Saeed Salem, Daehun Nyang, and David Mohaisen. Dl-fhmc: Deep learning-based fine-grained hierarchical learning approach for robust malware classification. *IEEE Transactions on Dependable and Secure Computing*, 2021.

[3] Hisham Alasmary, Ahmed Abusnaina, Rhongho Jang, Mohammed Abuhamad, Afsah Anwar, DaeHun Nyang, and David Mohaisen. Soteria: Detecting adversarial examples in control flow graph-based malware classifiers. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 888–898. IEEE, 2020.

[4] Firuz Juraev, Eldor Abdukhamidov, Mohammed Abuhamad, and Tamer Abuhmed. Depth, breadth, and complexity: Ways to attack and defend deep learning models. In *The 17th ACM ASIA Conference on Computer and Communications Security (ACM ASIACCS 2022)*, 2022.