School of Computing, Engineering, and the Built Environment
Edinburgh Napier University

PHD STUDENT PROJECT

Funding and application details

Funding status: Self-funded students only

Application instructions:
Detailed instructions are available at https://www.napier.ac.uk/research-and-innovation/research-degrees/how-to-apply

Prospective candidates are encouraged to contact the Director of Studies (see details below) to discuss the project and their suitability for it.

Project details

Supervisory Team:

- DIRECTOR OF STUDY: Dr Nikolaos Pitropakis (Email: N.Pitropakis@napier.ac.uk)
- 2ND SUPERVISOR: Prof William J Buchanan

Subject Group: Cyber-security and system engineering

Research Areas: Artificial Intelligence, Cyber Security, Machine Learning

Project Title: Hardening the Computing Continuum against Adversarial Attacks

Project description:

ML technologies rapidly expand as they perform predictions or support decision-making in healthcare, intrusion detection, fraud detection, autonomous vehicles, and many other applications. However, their popularity has rendered them attractive targets to adversaries who want to manipulate such mechanisms for malevolent purposes. Malicious actors can impact the decision-making algorithms of such approaches by either targeting the training data or forcing the model to their desired output, e.g., misclassification of abnormal events.

There is a variety of approaches proposed as defence mechanisms for the adversarial setting. They include defensive distillation, feature squeezing along with other approaches where the models are trained using adversarial examples. However, when malicious users change their methodology that creates adversarial examples, the ML mechanisms drop their accuracy again.

Attacks against Machine Learning can have a different background from the attacker's perspective (blackbox, greybox and whitebox attacks). The high-level goal of all attack models is to maximize the generalization error of the classification and possibly mislead the decision-making system towards desired malicious measurement values. The attacks are split into poisoning, where the adversary can poison the training dataset and evasion attacks where the adversary can undertake an evasion attack against classification during the testing phase thus producing a wrong system perception

The challenge is to create a robust methodology that addresses all the aforementioned attacks by acting:

a) Passively filtering adversarial examples; and
b) Actively retraining the model under adversarial settings.

The purpose of the Ph.D. project is:

a) to explore the literature with regard to attacks against machine learning;
b) design a methodology which will
    I. Detect adversarial attacks
    II. Mitigate the attacks

References:

[1] Pitropakis, Nikolaos, et al. "A taxonomy and survey of attacks against machine learning." Computer Science Review 34 (2019): 100199.
[2] Papadopoulos, Pavlos, et al. "Launching adversarial attacks against network intrusion detection systems for iot." Journal of Cybersecurity and Privacy 1.2 (2021): 252-273.
[3] Kantartopoulos, P., Pitropakis, N., Mylonas, A., & Kylilis, N. (2020). Exploring adversarial attacks and defences for fake twitter account detection. Technologies, 8(4), 64.
[4] Gallagher, Michael, et al. "Investigating machine learning attacks on financial time series models." Computers & Security 123 (2022): 102933.

Candidate characteristics

Education:

A second class honour degree or equivalent qualification in Computer Science, Computer Engineering, Data Science, Cyber Security, or Information Science

Subject knowledge:

- Machine Learning
- Coding skills in at least one object-oriented programming language
- Coding skills in at least one scripting programming language

Essential attributes:

- Curiosity and Inquisitiveness
- Self-Motivation and Discipline
- Adaptability and Resilience
- Critical Thinking and Analytical Skills

- Effective Communication

Desirable attributes: