Innovative Machine Learning Techniques for Cybersecurity

Academic supervisors	International supervisor	Thesis Evaluators
Supervisor: Annalisa Appice **Co-Supervisor:** Corrado Loglisci	**Supervisor:** Lorenzo Cavallaro	Battista Biggio Jerzy Stefanowski

As the number of interconnected devices increases, security attacks continue to advance at a rate outpacing cyber defenders' ability to write new attack signatures. In this scenario, Machine Learning (ML) plays a crucial role in modern malware and intrusion detection systems, due to its ability to detect unseen attacks. With the recent boom of Deep Learning (DL) in ML, the use of deep neural networks has emerged as a valuable candidate solution for various cybersecurity problems. However, the effectiveness of DL models is related to the capacity to discover malicious activities also under drifting conditions. Another challenge is to handle the imbalanced condition since a high-class imbalance is naturally inherent in the real-world cybersecurity domain. Finally, although DL models can make accurate decision, they perform as black-box models. So, it remains difficult to explain which characteristics of the input drive the decisions of a DL model.

Unseen attack detection

To improve the ability of DL techniques to detect unseen attacks in network traffic data, we investigate new DL approaches that try to reduce model overfitting during the learning stage. This is done by coupling local and global knowledge. We formulate two DL-based approaches that enrich the local description of a sample with global knowledge of normal and malicious behaviours synthesised through autoencoders trained on normal samples and attacks, respectively [1, 2]. In [3] we explore how to increase the ability of detecting unseen attacks by re-positioning the decision boundary that separates the training normal samples and the attacks before training the model described in [2]. In [4, 5], we describe two approaches that convert network traffic data in imagery and train Convolutional Neural Networks (CNNs) to delineate potential data patterns that appear on neighbour pixels once these pixels encode similar features of network flows. These two studies contribute to validating the idea that computer vision can aid in solving cybersecurity problems by accounting for the spatial continuity phenomenon among traffic characteristics.

Concept drift

To handle concept drift in network traffic data, we design two approaches [10] [11] that learn online a DL model on a network traffic data stream. In [10] we extend the framework proposed in [2] to handle a streaming scenario by combining Page Hinkley Test to detect concept drifts and transfer learning to fine-tune the DL model to the drifted data. In [11], we propose a semi-supervised intrusion detector that continuously updates the deep neural model as network traffic characteristics are affected by concept drift. We use active learning to reduce latency in the model updates, and label estimation to reduce labelling overhead.

Imbalanced data

Another contribution concerns the design of ML and DL approaches that allow us to deal with the imbalance of the malicious traffic. Because of the excellent performance of Generative Adversarial Networks (GANs) in image data augmentation, we investigate the viability of training GANs to augment images of network intrusions and balancing the training data set. The study in [5] shows that the image augmentation through GANs can be more effective than the application of traditional ML data augmentation techniques (e.g., SMOTE or ADASYN). However, generating artificial samples may cause overfitting, noise, or class overlap. So, we explore an alternative approach to deal with imbalanced data using deep metric learning (DML). In [6], we describe a Triplet network for network intrusion detection, which introduces a new triplet construction strategy that uses autoencoders to

derive both the positive and negative information for the triplet construction. The empirical study provides encouraging results, also in the multi-class scenario [7]. In addition, we explore ML solutions to handle the imbalanced issue also in Android malware detection tasks. To this aim, we couple knowledge extracted via clustering and classification, to deal with imbalanced data [8, 9].

Explainable Artificial Intelligence

Finally, we start the investigation of the transparency issue in DL-based algorithms by exploring the integration of XAI techniques in DL-based IDSs [11, 12]. In [11], we use XAI to interpret how the DNN model changes over the time to react on the drifting distribution in the processed network traffic data stream. In [12], we use XAI to produce visual explanations of traffic characteristics that more contribute to attack predictions and use these visual explanations to improve the accuracy of the IDS and explain how the decisions on the attacks are produced.

Part of this research work has been accomplished within a 4-months studentship at the Cybersecurity group's Systems Security Research Lab in the Department of Informatics at King's College London, under the supervision of Prof. Lorenzo Cavallaro.

References

[1] G. Andresini, A. Appice, N. D. Mauro, C. Loglisci e D. Malerba, «Exploiting the Auto-Encoder Residual Error for Intrusion Detection,» in *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS PW)*, 2019.

[2] G. Andresini, A. Appice, N. D. Mauro, C. Loglisci e D. Malerba, «Multi-Channel Deep Feature Learning for Intrusion Detection,» *IEEE Access,* vol. 8, pp. 53346-53359, 2020.

[3] G. Andresini, A. Appice, F. P. Caforio e D. Malerba, «Improving Cyber-Threat Detection by Moving the Boundary Around the Normal Samples,» in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*, vol. 919, 2021, pp. 105-127.

[4] G. Andresini, A. Appice e D. Malerba, «Nearest cluster-based intrusion detection through convolutional neural networks,» *Knowledge-Based Systems,* vol. 216, p. 106798, 2021.

[5] G. Andresini, A. Appice, L. D. Rose e D. Malerba, «GAN augmentation to deal with imbalance in imaging-based intrusion detection,» *Future Generation Computer Systems,* vol. 123, pp. 108-127, 2021.

[6] G. Andresini, A. Appice e D. Malerba, «Autoencoder-based Deep Metric Learning for Network Intrusion Detection,» *Information Sciences,* 2021.

[7] G. Andresini, A. Appice e D. Malerba, «A Two-Step Network Intrusion Detection System for Multi-Class Classification (Discussion Paper),» in *SEBD: 259-266 (2021)*, 2021.

[8] G. Andresini, A. Appice e D. Malerba, «Dealing with Class Imbalance in Android Malware Detection by Cascading Clustering and Classification,» in *Complex Pattern Mining - New Challenges, Methods and Applications*, vol. 880, Springer, 2020, p. 173–187.

[9] A. Appice, G. Andresini e D. Malerba, «Clustering-Aided Multi-View Classification: A Case Study on Android Malware Detection,» *J. Intell. Inf. Syst.,* vol. 55, p. 1–26, 2020.

[10] G. Andresini, A. Appice, C. Loglisci, V. Belvedere, D. Redavid e D. Malerba, «A Network Intrusion Detection System for Concept Drifting Network Traffic Data,» in *Discovery Science*, 2021.

[11] G. Andresini, F. Pendlebury, F. Pierazzi, C. Loglisci, A. Appice e L. Cavallaro, «INSOMNIA: Towards Concept-Drift Robustness in Network Intrusion Detection,» in *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2021.

[12] F. P. Caforio, G. Andresini, G. Vessio, A. Appice e D. Malerba, «Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems,» in *Discovery Science*, 2021.