Αλγόριθμοι Παραγωγικής Τεχνητής Νοημοσύνης (Generative Artificial Intelligence - GenAI) για την Ερμηνεία των Αποτελεσμάτων Μεθόδων eXplainable AI (XAI)

Η χρήση τεχνικών eXplainable Artificial Intelligence (XAI) για την ερμηνεία των αποτελεσμάτων μοντέλων μηχανικής μάθησης έχει αυξηθεί ραγδαία στο πεδίο της ασφάλειας δικτύων υπολογιστών [1, 2]. Τα πολυάριθμα είδη διαγραμμάτων που παράγονται από τις μεθόδους XAI βοηθούν τους χρήστες και τους προγραμματιστές των μοντέλων μηχανικής μάθησης να κατανοήσουν ποια χαρακτηριστικά (features) είναι τα πιο επιδραστικά, καθώς και πώς συγκεκριμένες τιμές επηρεάζουν τα αποτελέσματα των μοντέλων [2].

Ωστόσο, τα διαγράμματα που επιστρέφονται από τις μεθόδους XAI είναι συνήθως περίπλοκα και απαιτούν την καταβολή ιδιαίτερης προσπάθειας για τη σωστή κατανόησή τους. Μία ιδιαίτερα υποσχόμενη κατεύθυνση είναι η χρήση των αλγορίθμων παραγωγικής τεχνητής νοημοσύνης (Generative Artificial Intelligence - GenAI) και των Large Language Models (LLM's) για την αυτόματη παραγωγή επεξηγήσεων από διαγράμματα που επιστρέφονται ως έξοδοι από τις τεχνικές XAI [3].

Η συγκεκριμένη διπλωματική θα μελετήσει αρχικά τη χρήση μεθόδων XAI για την παραγωγή και την εμφάνιση επεξηγήσεων σε προβλήματα ασφάλειας δικτύων υπολογιστών. Μέθοδοι που μπορούν να εξεταστούν είναι η SHapley Additive exPlanations (SHAP) και η Local Interpretable Model-agnostic Explanations (LIME) [4], ενώ προβλήματα ασφάλειας δικτύων που μπορούν να μελετηθούν είναι η ανίχνευση ονομάτων που παράγονται από Domain Generation Algorithms (DGA's) [5] ή η ανίχνευση επιθέσεων Distributed Denial of Service (DDoS) [6]. Στη συνέχεια της διπλωματικής, θα χρησιμοποιηθούν εργαλεία GenAI και LLM's, π.χ. ChatGPT [7] ή ανοιχτά γλωσσικά μοντέλα όπως το Llama [8], για την παραγωγή κειμένου σε φυσική γλώσσα που θα περιγράφει με απλουστευμένο τρόπο τα αποτελέσματα των εξόδων αλγορίθμων XAI.

Η διπλωματική θα βασιστεί σε σύνολα δεδομένων (datasets), που χρησιμοποιούνται ευρέως από ερευνητές που δραστηριοποιούνται στο πεδίο της ασφάλειας δικτύων υπολογιστών [5, 9].

[1] T. Zebin, S. Rezvy and Y. Luo, "An Explainable AI-based Intrusion Detection System for DNS over HTTPS (DoH) Attacks", IEEE Transactions on Information Forensics and Security, Volume 17, March 2022
[2] N. Kostopoulos, D. Kalogeras, D. Pantazatos, M. Grammatikou and V. Maglaris, "SHAP interpretations of tree and neural network DNS classifiers for analyzing DGA family characteristics", IEEE Access, Volume 11, pp. 61144-61160, June 2023
[3] C.C. Hsu, I.Z. Wu and S.M. Liu, "Decoding AI Complexity: SHAP Textual Explanations via LLM for Improved Model Transparency", 2024 International IEEE Conference on Consumer Electronics, pp. 197-198, July 2024
[4] Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/
[5] D. Plohmann, K. Yakdan, M. Klatt, J. Bader and E. Gerhards-Padilla, "A comprehensive measurement study of domain generating malware", USENIX Security Symposium, pp. 263–278, August 2016
[6] S.T. Zargar, J. Joshi and D. Tipper, "A Survey of Defense Mechanisms against Distributed Denial of Service (DDoS) Flooding Attacks", IEEE Communications Surveys & Tutorials, Volume 15, pp. 2046-2069, March 2013
[7] ChatGPT, https://openai.com/chatgpt/
[8] Llama, https://www.llama.com/
[9] CIC-DDoS2019, https://www.unb.ca/cic/datasets/ddos-2019.html