



RISE

Infrastruttura Strategica della Ricerca
UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Intelligenza artificiale spiegabile per la scoperta delle cyber-minacce



Per informazioni rivolgersi alle responsabili scientifiche:

Prof.ssa Annalisa Appice
annalisa.appice@uniba.it

Dr Giuseppina Andresini
giuseppina.andresini@uniba.it

Polo Scientifico Tecnologico

Dipartimento di Informatica

Via Orabona 4, 70125 (BA)

ENG: Within Modern cybersecurity systems are designed to promptly recognize cyber threats, such as malware and network intrusions, by preventing attackers from having enough time to gain access to sensitive data or services. Over the past decades, Artificial Intelligence (AI) and, in particular, Deep Learning (DL) methods have contributed to make traditional, signature-based, cybersecurity systems obsolete, thanks to their ability to learn complex relationships from cyber-data, by gaining accuracy in detecting ever-evolving cyber threats. Although the priority of modern cybersecurity systems remains the accurate detection and classification of cyber-threats, the explainability of these detection models has recently emerged as a further important requirement for the integration of such systems in commercial tools. Explaining cyber threat alerts can help increase the trustiness of security analysts in AI-powered detection systems by also disclosing useful insights to handle the malicious behaviour of detected threats.

Within the PNRR projects FAIR – Future AI Research (PE00000013) and SERICS (PE00000014), both funded by the European Union through NextGenerationEU, the University of Bari Aldo Moro developed an integrated service that leverages Deep Learning (DL) methods, appropriately enhanced with eXplainable AI (XAI) facilities, to automate the identification and explanation of cyber threats in both network traffic [Al-Essa et al. 2024; De Rose et al. 2024] and Android devices [Al-Essa et al. 2024; De Rose et al. 2024; Andresini et al. 2026]. The methodological components of the service include sophisticated DL architectures (e.g., Vision Transformer and Graph Neural Network) combined with knowledge distillation and ensemble learning strategies. These components are equipped with explainability facilities that use explanations for ensemble model selection or knowledge distillation [De Rose et al. 2024] by improving predictive accuracy of the detection phase, as well as extracting explanations of cyber-threat signatures [Al-Essa et al. 2024; De Rose et al. 2024; Andresini et al. 2026] to describe the signature of correctly detected malware and perform the root cause analysis of false malware alerts [Andresini et al. 2026].

The goal is to provide an integrated service to allow Security Analysts to take advantage of accurate cyber threat predictions equipped with explanations in a Symbiotic AI perspective. The demo code of the service is freely available on GitHub:

<https://github.com/malikalessa/PANACEA>

<https://github.com/Kyanji/VINCENT>

<https://github.com/TheBlakeOmega/ANAKIN>

ITA: I moderni sistemi di sicurezza informatica sono progettati per riconoscere tempestivamente minacce informatiche (cyber-minacce), quali malware e intrusioni di rete, riducendo il rischio che gli attaccanti dispongano del tempo necessario per accedere a dati o servizi sensibili. Negli ultimi decenni, l'Intelligenza Artificiale (IA) e, in particolare, i metodi di Deep Learning (DL) hanno reso obsoleti i sistemi di sicurezza tradizionali basati su firme, grazie alla loro capacità di apprendere relazioni complesse da cyber-dati e migliorare l'accuratezza del rilevamento di cyber-minacce in continua evoluzione. Sebbene la priorità dei moderni sistemi di sicurezza rimanga il rilevamento e la classificazione accurati delle cyber-minacce, la spiegabilità di tali modelli è recentemente emersa come un ulteriore requisito fondamentale per la loro effettiva integrazione in strumenti commerciali. Spiegare gli “alert” relativi alle cyber-minacce può infatti accrescere la fiducia degli analisti della sicurezza nei sistemi di rilevamento basati sull'IA, fornendo al contempo informazioni utili per gestire i comportamenti malevoli delle minacce individuate. Sulla base di queste premesse, nell'ambito dei

progetti PNRR FAIR – Future AI Research (PE00000013) e SERICS (PE00000014) – entrambi finanziati dall'Unione Europea tramite NextGenerationEU – l'Università di Bari Aldo Moro ha sviluppato un servizio integrato che trae vantaggio dall'uso di metodi di Deep Learning (DL), opportunamente potenziati con funzionalità di eXplainable AI – XAI (IA spiegabile), per automatizzare l'identificazione e la spiegazione di cyber-minacce sia nel monitoraggio del traffico di rete [Al-Essa et al. 2024; De Rose et al. 2024] sia nel monitoraggio di applicazioni installate su dispositivi Android [Al-Essa et al. 2024; De Rose et al. 2024; Andresini et al. 2026]. Le componenti metodologiche del servizio includono sofisticate architetture DL (come Vision Transformer e Graph Neural Network) combinate con strategie di “knowledge distillation” ed “ensemble learning”. Tali componenti sono dotate di funzionalità di spiegabilità che utilizzano le spiegazioni per la selezione del modello “ensemble” o per completare la “knowledge distillation” di un sistema di rilevamento [De Rose et al. 2024]. Tali componenti migliorano l'accuratezza predittiva della fase di rilevamento, e consentono di estrarre firme delle minacce informatiche [Al-Essa et al. 2024; De Rose et al. 2024; Andresini et al. 2026] per descrivere le caratteristiche dei comportamenti malevoli ed eseguire l'analisi delle cause (“root cause analysis”) in caso di falsi positivi relativi a malware [Andresini et al. 2026].

L'obiettivo è mettere a disposizioni degli analisti della sicurezza un servizio integrato che colleghi previsioni accurate sulle minacce informatiche a spiegazioni dettagliate, guidando così le decisioni atte a tutelare la sicurezza delle infrastrutture digitali in un'ottica di IA Simbiotica (Symbiotic AI). Il codice demo dimostrativo del servizio è disponibile sui GitHub:

<https://github.com/malikalessa/PANACEA>

<https://github.com/Kyanji/VINCENT>

<https://github.com/TheBlakeOmega/ANAKIN>

References

- AL-Essa, M., Andresini, G., Appice, A. et al. PANACEA: a neural model ensemble for cyber-threat detection. Mach Learn 113, 5379–5422 (2024). <https://doi.org/10.1007/s10994-023-06470-2>
- [De Rose L., Andresini G., Appice A., Malerba D. , VINCENT: Cyber-threat detection through vision transformers and knowledge distillation, Computers & Security, Volume 144, 2024, 103926, <https://doi.org/10.1016/j.cose.2024.103926>.

- Andresini, G., Appice, A., Belvedere, V. et al. Anakin: explainable android malware detection with graph neural networks. *Cybersecurity* 9, 116 (2026). <https://doi.org/10.1186/s42400-026-00552-z>