PhD student Andrea Vignali

Improving security of networked systems through an NLP-based Anomaly Detection approach

Tutor: G. Sperlì

Cycle: XXXVIII

co-Tutor: S.P. Romano

Year: Second

My background

- MSc degree in Computer Engineering @ DIETI – Federico II
  - Thesis: “An active learning and similarity based augmentation approach for few-shot NER applications”
- Research group/laboratory: PICUSlab and ARCLab
- PhD start date: November 1, 2022
- Scholarship type: PNRR – DM 352
- Partner company:
  - AKKODIS ITALY S.R.L. (former AKKA Italia s.r.l.)
- Period in company: November 1, 2023 – May 1, 2024
- Period abroad: September 2, 2024 – Ongoing (February 28, 2025) @ Massachusetts Institute of Technology – CSAIL – ALFA

Summary of study activities

- 2 Ad hoc courses + 1 PhD school
  - Hands-on Network Intrusion Detection via Machine and Deep Learning
  - Strategic Orientation for STEM Research & Writing
  - Open and programmable 6G networks in the cloud/edge continuum: research challenges and experimentation tools in SLICES Research Infrastructures
- 9 Seminars
- Conference
  - 1st International Workshop On Signal Processing For Resilient Intrusion Detection In Cyber-Physical Systems SPID-CPS 2024 @ 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)
- Period in company: 6 months (November 1, 2023 – May 1, 2024)
- Period abroad: September 2, 2024 – Ongoing @ Massachusetts Institute of Technology – CSAIL – ALFA

Research field of interest

- **NLP and Anomaly Detection applied to Cybersecurity**

- **Natural Language Processing (NLP)**
  - **NLP_RQ1:** How can we address challenges related to data scarcity in NLP tasks?
  - **NLP_RQ2:** How can models use semi-structured and unstructured text?
  - **NLP_RQ3:** How LLMs and other models are used in other domains?

- **Anomaly Detection**
  - **AD_RQ1:** What are effective strategies for handling multiple data sources in anomaly detection?
  - **AD_RQ2:** How can anomaly detection techniques be applied effectively within industrial settings?

Research results (NLP)

- **NLP_RQ1:** Data augmentation techniques [P1], including reinforcement learning [S1] and active learning [A1], have proven effective in enhancing NLP models. In particular, Large Language Models (LLMs) stand out in general domain of application.
- **NLP_RQ2:** When prompted effectively, LLMs can interpret semi-structured text, enabling them to generate code [S8] and extract sentiment features [S7].
- **NLP_RQ3:** NLP models offer valuable applications across fields with diverse or unstructured/semi-structured textual data sources, such as Finance [P3, S7], Law, and Software Testing[S8].

Research results (Anomaly Detection)

- **AD_RQ1:** Using tailored models for each data source, following a hyperparameter tuning phase, has proven effective [S2]. Additionally, late fusion techniques [S3] and the integration of graph neural networks with time series correlation [S5] have shown to be valuable for managing multiple data sources.
- **AD_RQ2:** Anomaly detection models have demonstrated strong performance in pump fault detection [P2]. When combined with NLP techniques to generate machine-readable representations of text, these models have also been effective in test prioritization [S8].

Research products

[P1]	***Named Entity Recognition using context similarity data augmentation*** *– Ilaria Bartolini, Angelo Chianese, Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, Andrea Vignali – conference: 32nd Symposium on Advanced Database Systems (SEBD 2024) –* ***Published*** *– 2024*
[P2]	***Anomaly Detection in Cyber-Physical Systems: A Case Study on Pump Health Monitoring*** *– Giancarlo Sperlì, Andrea Vignali – conference: 1st International Workshop On Signal Processing For Resilient Intrusion Detection In Cyber-Physical Systems SPID-CPS 2024 @ 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW) –* ***Published*** *– 2024*
[P3]	***An NLP-based approach to assessing a company's maturity level in the digital era*** *– Simon Pietro Romano, Giancarlo Sperlì, Andrea Vignali – journal: Expert Systems with Applications –* ***Published*** *– 2024*
[A1]	***ALDANER: Active Learning based Data Augmentation for Named Entity Recognition*** *– Vincenzo Moscato, Marco Postiglione, Giancarlo Sperlì, Andrea Vignali – journal: Knowledge-Based Systems –* ***Accepted*** *– 2024*

Research products

[S1]	***PALAUNER: Policy-based Active Learning to AUgment Named Entity Recognition datasets*** *– Marco Postiglione, Andrea Vignali, Giancarlo Sperlì, Guido Maria Secondulfo, Vincenzo Moscato – journal: Neural Computing and Applications –* ***Submitted*** *– 2023*
[S2]	***Empowered Cyber-Physical Systems Security using both Network and Physical Data*** *– Roberto Canonico, Giovanni Esposito, Annalisa Navarro, Simon Pietro Romano, Giancarlo Sperlì, Andrea Vignali – journal: Computers & Security –* ***Submitted*** *– 2024*
[S3]	***An Anomaly-based Approach for Cyber-Physical Threat Detection using Network and Sensor Data*** *– Roberto Canonico, Giovanni Esposito, Annalisa Navarro, Simon Pietro Romano, Giancarlo Sperlì, Andrea Vignali – journal: Computer Communications –* ***Submitted*** *– 2024*
[S4]	***Smart home demand-side management based on rooftop deep learning photovoltaic power forecasting*** *– Pasquale De Falco, Giancarlo Sperlì, Marcello Vestri, Andrea Vignali – journal: Journal of Network and Computer Applications –* ***Submitted*** *– 2024*

Research products

[S5]	***Threat detection in reconfigurable Cyber-Physical Systems through Spatio-Temporal Anomaly Detection using Graph Attention Network*** *– Roberto Canonico, Francesco Lista, Annalisa Navarro, Giancarlo Sperlì, Andrea Vignali – journal: Engineering Applications of Artificial Intelligence –* ***Submitted*** *– 2024*
[S6]	***Empowering Code Translation with Large Language Models integrating Human-in-the-Loop Feedback*** *– Gabriele Dario De Siano, Anna Rita Fasolino, Giancarlo Sperlì, Andrea Vignali – journal: Information and Software Technology –* ***Submitted*** *– 2024*
[S7]	***Leveraging Large Language Models for Sentiment-Driven Stock Market Forecasting*** *– Giovanni Officioso, Andrea Vignali – conference: Artificial Intelligence for Financial Domain @ ACM/SIGAPP Symposium On Applied Computing (AIFD@SAC 2025) –* ***Submitted*** *– 2024*
[S8]	***Harnessing NLP for intelligent testing: unsupervised approaches*** *– Andrea Vignali, Giancarlo Sperlì, Simon Pietro Romano – conference: Industry track @ IEEE International Conference on Software Testing, Verification and Validation (ICST 2025) –* ***Submitted*** *– 2024*

Future work

- Currently working on an AI-driven solution for Automatic Exploit Chain Discovery within cybersecurity.
- **Problem:** Automating the discovery of exploit chains based on the structure and vulnerabilities of a given network configuration.
- **Challenges:**
  - Multiple unstructured and semi-structured data sources
  - Scalability of the system
  - Model Adaptability and Robustness
- **Methodology:**
  - LLM for Data Interpretation
  - Planning Domain Definition Language translation of the data and Domain Modeling with a parser
  - AI Planning to solve a PDDL-formatted exploit chain discovery problem