Course Title

Understanding symbiotic AI systems to improve robustness of models

Teacher(s)

Giuseppina Andresini

Course Website (optional)

Code: To be defined

Course description (min 150, max 300 words)

In the rapidly evolving landscape of Artificial Intelligence (AI), the need for robust and reliable AI models is more critical than ever. As AI continues to play an increasing role in our daily lives, it has become imperative to ensure that AI models can consistently perform well and make trustworthy decisions. The Algorithmic Impact Assessment is a critical process that evaluates the impact of AI systems from an ethical and human perspective by focusing on Explainability, Fairness and Robustness points of view. One key aspect that significantly contributes to the understandability and trustworthiness of symbiotic AI systems is the provision of explanations of decision models. These explanations serve as the bridge between complex AI algorithms and human understanding. Research and practical experience have proved that when explanations are provided alongside AI system outputs, users typically increase their trust in algorithms.

In this course, we will present an overview of Explainable Artificial Intelligence (XAI) strategies to produce explanations of the behaviour of algorithms and justification of model decisions (e.g., global or local algorithms, agnostic or specific algorithms). In particular, we will showcase several examples of recent progress in applying XAI in order to account for specific critical conditions (e.g., imbalanced data, concept drifts, adversarial samples), as well as methods to account for knowledge disclosed with explanations, in order to improve robustness of symbiotic models (e.g., distillation, defensive distillation, attention). Specifically, the course will be organized into two main sections:
(i) Generating explanations and decision justifications: this section is likely to cover the strategies and techniques for creating understandable explanations.
(ii) Learning robust AI models accounting for explanations: this section will likely focus on methods and practices for building more robust AI models including model reliability, interpretability, and resilience to adversarial attacks.

Course period
November-December 2023

Course References (optional)

[1] Leonida Gianfagna , Antonio Di Cecco, Explainable AI with Python, Springer Cham (2021) ISSN 978-3-030-68639-0, https://doi.org/10.1007/978-3-030-68640-6 .

[2] Denis Rothman, Sotiris Moschoyiannis, Helge Janicke,
Hands-On Explainable AI (XAI) with Python, Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps, Packt, 2020, 102419, ISSN 9781800208131.

[3] Zixiao Kong, Jingfeng Xue, Yong Wang, Lu Huang, Zequn Niu, Feng Li, A Survey on Adversarial Attack in the Age of Artificial Intelligence, Wireless Communications and Mobile Computing, vol. 2021, Article ID 4907754, 22 pages, 2021. https://doi.org/10.1155/2021/4907754

Credits and Hours

3 credits, two of lectures (16 hours) and one of practice (15 hours), for a total of 31 hours.

Exam Modality

Two alternatives are available to the student to pass this exam:

1) Paper presentation. Students will illustrate 1 paper suggested by the teachers by discussing novelties and issues and identifying possible relationships with their research projects. No groups are allowed
2) Project. Students will implement and experimentally validate an XAI-based approach for an ML technique suggested by the teachers. Projects can be done in groups of 1-3 students, depending on the complexity of the technique.

Teacher(s) CV

Giuseppina Andresini is an Assistant Professor (non-tenure) at the Department of Computer Science, the University of Bari Aldo Moro. She graduated cum Laude in Computer Science from the University of Bari Aldo Moro, Italy, in April 2018, discussing a Laurea thesis on data mining. She was a Ph.D. Student in Computer Science (with a scholarship) from November 2, 2018 to November 1, 2021 with a research project on “Innovative machine learning techniques for cybersecurity”. She received a Ph.D. in Computer Science, University of Bari Aldo Moro, Italy, in March 2022. Her current research interests mainly concern deep learning, XAI and adversarial learning with applications in cybersecurity and remote sensing. On these topics, she has published several papers in international journals and international conferences. She has been a member of the Organization Committee of the CyberChallenge.IT 2020 and 2021 in the Department of Computer Science, University of Bari Aldo Moro. She has participated in the organization (as co-chair) of the three editions of Workshop on Machine Learning for Cybersecurity co-located with ECML PKDD 2021-2023. She held 2 CFU (16 hours) of the course Advanced Machine Learning for Cyber Defense at PhD Programme in Computer Science and Mathematics, University of Bari, Italy - Cycle XXXVII. She held the module of Deep Learning Advanced (12 hours) for the course of

Machine Learning Techniques and Applications of Master 2 DISS (Data Intelligence for Smart Systems) at Lyon 1 University (Lyon, France). She has co-tutored 20 theses in both Bachelor and Master degrees on topics cybersecurity, machine learning, adversarial learning and deep learning.

Teacher(s) Main Publications

1. AL-Essa M., Andresini G., Appice A., Malerba D., **PANACEA: A Neural Model Ensemble for Cyber-Threat Detection,** Accepted for publication in Machine Learning Journal for Journal Track of Emerging Applications and Frontiers for Data Science (DSAA 2023)
2. **Andresini G.**, Appice A**,** Malerba D., **SILVIA: An eXplainable Framework to Map Bark Beetle Infestation in Sentinel-2 Images,** IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, Pages 1-17, OI: 10.1109/JSTARS.2023.3312521, *ad oggi, City Score 7.8 IF 5.5*
3. **Andresini G.**, Appice A**,** Ienco D.., Malerba D., **Seneca: Change detection in optical imagery using siamese networks with active-transfer learning,** Expert Systems with Applications, Volume 214, 2023, Pages 108-127 ISSN 09574174, DOI: 10.1016/j.eswa.2022.119123, *ad oggi, City Score 12.2 IF 8.665*
4. **Andresini G.**, Appice A**,** Caforio F., Malerba D., Vessio G. **Roulette: A neural attention multi-output model for explainable network intrusion detection,** Expert Systems with Applications, Volume 214, 2022, Pages 108-127 ISSN 09574174, DOI:10.1016/j.eswa.2022.117144, *ad oggi, City Score 12.2 IF 8.665*
5. **Andresini G.**, Appice A**,** De Rose L., Malerba D., **GAN augmentation to deal with imbalance in imaging-based intrusion detection**, *Future Generation Computer Systems, Elsevier*, Volume 123, 2021, Pages 108-127 ISSN 0167-739X, DOI: 10.1016/j.future.2021.04.017, *ad oggi, City Score 13.3 IF 7.187*
6. **Andresini G.**, Appice A., Malerba D., **Autoencoder-based deep metric learning for network intrusion detection**, *Information Sciences, Elsevier*, Volume 569, 2021, Pages 706-727, ISSN 0020-0255 DOI: 10.1016/j.ins.2021.05.016*, ad oggi, City Score 12.1 IF 6.795*
7. **Andresini G.**, Appice A., Malerba D., **Nearest cluster-based intrusion detection through convolutional neural networks**, *Knowledge-Based Systems, Elsevier,* Volume 216, 2021,106798, ISSN 0950-7051, DOI: 10.1016/j.knosys.2021.106798, *ad oggi, City Score 11.3 IF 8.038*
8. **Appice A, Andresini G.**, Malerba D., **Clustering-Aided Multi-View Classification: A Case Study on Android Malware Detection**. *J Intell Inf Syst, Springer,* 55, 1–26 (2020). DOI: 10.1007/s10844-020-00598-6, *ad oggi, City Score 4.4 IF 1.83*
9. **Andresini, G**., Pendlebury, F., Pierazzi, F., Loglisci, C., **Appice,** A., Cavallaro, L. **INSOMNIA: Towards Concept-Drift Robustness in Network Intrusion Detection**. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security (AISec)* ACM.
10. Caforio F.P., **Andresini G.**, Vessio G., Appice A., Malerba D,. **Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems**. Discovery Science. DS 2021. Lecture Notes in Computer Science(), vol 12986. Springer, Cham. DOI: 0.1007/978-3-030-88942-5_30