Doctoral School of Information and Biomedical Technologies
Polish Academy of Sciences (TIB PAN)

SUBJECT:
AI Security - Adversarial Attacks on Deepfake Detection Models and Design of Novel Detection Models

SUPERVISOR:

- prof. dr. hab. Franck Leprévost (Primary co-supervisor)
  University of Luxembourg
  franck.leprevost@uni.lu
- dr. hab. Joanna Kołodziej (prof. NASK, Co-supervisor)
  Naukowa i Akademicka Sieć Komputerowa (NASK)
  joanna.kolodzieg@nask.pl

DESCRIPTION:
Recent progress in generative AI has made it easy to create very realistic fake images, videos and audio. While this technology can be useful for some applications, it also raises risks, as it can be misused to spread false information or distort real events. Because of these concerns, one current research direction of AI security is to develop better methods to detect deepfakes in visual and audio domains.

On the one hand, based on manipulation level, visual deepfakes are grouped into different types like face swap/ identity swap, lip-syncing, face-reenactment/ puppet-mastery, entire face synthesis and facial attribute manipulation, while audio deepfakes are further classified as text-to-speech synthesis and voice conversion.

On the other hand, deepfake detection methods often rely on identifying artifacts and inconsistencies left during the generation process (inconsistencies in head pose, lack of eye blinking, color variations in facial texture, teeth alignment, spatial-temporal features, psychological signals like heart rate, individual's behavior patterns, etc.).

However, these methods are vulnerable to evasion techniques that aim to remove or disguise such traces. Attackers can employ several strategies to deceive detectors, including adversarial perturbations (e.g., random cropping, noise, or JPEG compression), elimination of manipulation traces in the frequency domain (e.g. by enhancing spectral distributions of GAN-generated samples), and advanced image filtering (to erase generation fingerprints or introduce misleading noise). Prior studies

have shown that such attacks can significantly reduce the accuracy of state-of-the-art deepfake detectors in both visual and audio domains.

This PhD research will focus on systematically attacking deepfake detection models using different evasion techniques to evaluate their robustness and uncover critical weaknesses. Building on the insights gained from these attacks, new deepfake detection models that are more resistant to adversarial perturbations will be developed.

REQUIREMENTS:

- MSc degree in computer science, mathematics, AI, or related field
- High programming skills in Python
- Experience with High Performance Computing (HPC)
- Experience with Computer Vision and image classifiers (CNNs and transformers)
- Experience with Automatic Speech Recognition (ASR) models
- Ability to conduct independent research and collaborate in an interdisciplinary team.
- Publication in Scientific Journal is a plus.
- Advanced level of English (spoken and written).

BIBLIOGRAPHY:

- Heidari, Arash, et al. "Deepfake detection using deep learning methods: A systematic and comprehensive review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 14.2 (2024): e1520.
- Carlini, Nicholas, and Hany Farid. "Evading deepfake-image detectors with white- and black-box attacks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020.
- Rabhi, Mouna, Spiridon Bakiras, and Roberto Di Pietro. "Audio-deepfake detection: Adversarial attacks and countermeasures." *Expert Systems with Applications* 250 (2024): 123941.
- Gowrisankar, Balachandar, and Vrizlynn LL Thing. "An adversarial attack approach for eXplainable AI evaluation on deepfake detection models." *Computers & Security* 139 (2024): 103684.
- Masood, Momina, et al. "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward." *Applied intelligence* 53.4 (2023): 3974-4026.
- Agarwal, Shruti, et al. "Protecting world leaders against deep fakes." *CVPR workshops*. Vol. 1. No. 38. 2019.