Contribution ID: **12** Type: **not specified**

Enhancing Nuclear Security Against AI Threats Through an Efficient Digital Forgery Detection Scheme

Digital threats such as deepfakes and digital forgery have become significant challenges due to the rapid advancements in Artificial Intelligence (AI). This is particularly in the vital sectors like nuclear security where sensitive data, images, and documents are involved. The manipulation of digital content poses risks to the integrity of nuclear operations and inspections. This objective of this paper is to develop an efficient scheme that overcomes advanced AI threats by detecting forgery in digital images and documents to enhance nuclear security and eliminate the AI-driven threats. The proposed scheme utilizes Vision Transformer (ViT) to capture intricate spatial patterns and inconsistencies within images for spatial feature extraction. Then, the detection framework is developed using transformer-based models for classifying the tested images into manipulated forged image or authentic image. Unlike traditional convolutional neural networks (CNNs), ViT segments the input images into fixed size patches then applies self attention algorithm to learn contextual relationships across entire images. This enables the model to detect spatial inconsistencies indicative of AI-generated forgery that might evade conventional detection techniques. The proposed scheme verified through several authentic samples of nuclear related images and documents including inspection images or official documents related to nuclear operations. Then generated synthetic forgeries using Generative Adversarial Networks (GANs) and different AI tools for simulating real threats. Structural similarity (SSIM) has been used for measuring the similarity between the tested image features and the stored features that feed the classifier to determine the classification accuracy of the proposed scheme. Performance evaluation was carried out using multiple different metrics, including classification accuracy, precision, recall, F1-score. The proposed scheme has been tested against several AI attacks such as deepfakes, inpainting & outpointing AI image editing, neural style transfer, morphing, attribute manipulation, image-to-image translation, and AI retouching and enhancement. The testing results proved the superiority of the proposed scheme, achieving an average detection accuracy above 98.5% for most of the manipulation treats, outperforming several benchmark models. Moreover, while minor performance degradation was observed under certain AI attacks, particularly subtle inpainting and neural style transfers, the proposed method maintained reliable detection capability, illustrating its practical applicability in nuclear security contexts. By combining the strengths of Vision Transformers, attention-driven classifiers, and structural similarity analysis, the proposed method offers a robust solution capable of enhancing nuclear operational security, safeguarding sensitive digital content, and contributing to the broader field of AI-driven cybersecurity mechanisms.

Author: MAHMOUD, Hani (Nuclear Research Center, Atomic Energy Authority, Egypt)

Co-author: Dr ELHADY, Walla (Sadat Academy for Management Sciences , Egypt)

Presenter: MAHMOUD, Hani (Nuclear Research Center, Atomic Energy Authority, Egypt)