ID 193

Systematic Review on Analyzing the Most Effective Method for Deepfake Detection of Images Generated by AI

DMUC Senarathne[1#] and BVKI Vidanage[2]

[1,2]Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University

[#]39-bcs-0018@kdu.ac.lk

Abstract

The rapid advancements in Artificial Intelligence and Deep Learning have revolutionized synthetic media creation, with Deepfakes emerging as a significant societal threat. These manipulated images and videos, often generated using Generative Adversarial Networks (GANs) and Diffusion Models (DMs), now turn out increasingly realistic due to modern technologies, making their detection more challenging and critical. Deepfakes cause serious risks to privacy, cybersecurity, and the spread of misinformation, especially across platforms like social media, where their rapid dissemination serves to undermine trust and public discourse. This systematic review, conducted in line with the PRISMA framework, examines over 50 research papers on Deepfake detection methods, focused on Convolutional Neural Networks (CNNs), frequency-based approaches, and hybrid models. The findings admit that CNNs excel in controlled settings, providing high detection accuracy, but poor generalization when applied to diverse and evolving Deepfake datasets. Hybrid models, while more adaptable to new manipulations, face significant limitations due to high computational costs, impeding their scalability for real-time applications. This study underscores the critical need for robust and scalable detection systems that are capable of performing well in real-time to support applications like social media moderation, cybersecurity defenses, and misinformation prevention. The insights aim to guide future research toward developing versatile and high-accuracy detection frameworks to neutralize the escalating threats posed by increasingly realistic Deepfakes.

Keywords: *Deepfake detection, AI-generated images, Diffusion models, Convolutional neural networks, Hybrid models*