SEEING THROUGH DEEPFAKES: AN EXPLAINABLE MULTI-CNN FRAMEWORK FOR DEEPFAKE DETECTION

Author's: Aasha A[1], R. Surendiran[2], K. Sujith[3], N. Vanjulavalli[4]

The proliferation of deepfake technology has posed significant challenges to digital media integrity, social trust, and cybersecurity. Deepfakes leverage advanced generative models to manipulate or synthesize human faces in images and videos, creating content that is visually convincing yet misleading or malicious. Detecting such manipulations is crucial for applications ranging from social media moderation and forensic investigations to privacy protection and political security. Traditional deepfake detection methods often rely on single convolutional neural network (CNN) models, which can be limited in their ability to generalize across diverse manipulations and datasets. This research proposes a robust, Seeing Through Deepfakes: An Explainable Multi-CNN Framework for Deepfake Detection to enhance deepfake detection accuracy, interpretability, and adaptability. The proposed methodology integrates multiple CNN architectures—including a lightweight baseline CNN, a deeper CNN model, and a residual CNN variant—to form an ensemble capable of capturing both low-level texture inconsistencies and high-level semantic cues in facial imagery. By leveraging the complementary strengths of each network, the ensemble model improves robustness against a wide variety of deepfake generation techniques. Synthetic datasets are employed for initial model development, simulating realistic artifacts such as ed ge irregularities, periodic texture anomalies, and noise patterns commonly present in manipulated images. This enables efficient experimentation while ensuring controlled evaluation metrics. To address the critical need for transparency in AI-driven detection, the framework incorporates Grad-CAM-based explainability.