ABSTRAK

Teknologi *deepfake* memungkinkan manipulasi gambar dan audio yang nyaris tidak terdeteksi, menimbulkan ancaman seperti penipuan daring, disinformasi, dan pelanggaran privasi. Penelitian ini bertujuan merancang sistem deteksi manipulasi *deepfake* multimodal menggunakan integrasi *Convolutional Neural Networks* (CNN) dan *Capsule Networks* (CapsNet) dengann interpretasi *Explainable Artificial Intelligence* (XAI). CNN mengekstrak fitur visual seperti tekstur dan gradien, sedangkan CapsNet menangkap hubungan spasial dan temporal antar fitur. Teknik XAI, seperti *Gradient-weighted Class Activation Mapping* (Grad-CAM) dan analisis fitur akustik (*Mel-Frequency Cepstral Coefficients*), memvisualisasikan area kritis pada gambar (mata, hidung, mulut) dan anomali temporal pada audio. Pengujian pada dataset publik dan data simulasi menghasilkan akurasi 80,80% (audio) dan 80,75% (gambar), dengan presisi 84,80%, *recall* 78,50%, dan F1-score 81,50% pada *batch size* 16. Penelitian ini berkontribusi pada pengembangan metode deteksi *deepfake* yang transparan secara akademik dan mendukung keamanan informasi secara aplikatif. Namun, peningkatan akurasi dan ketahanan terhadap manipulasi canggih masih diperlukan.

Kata Kunci: *Deepfake, Deep Learning, Convolutional Neural Networks, Capsule Networks, Explainable AI.*

ABSTRACT

Deepfake technology enables near-undetectable image and audio manipulation, posing threats such as online fraud, disinformation, and privacy violations. This research aims to design a multimodal deepfake manipulation detection system using the integration of Convolutional Neural Networks (CNN) and Capsule Networks (CapsNet) with Explainable Artificial Intelligence (XAI) interpretation. CNN extracts visual features such as texture and gradient, while CapsNet captures spatial and temporal relationships between features. XAI techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM) and acoustic feature analysis (Mel-Frequency Cepstral Coefficients), visualize critical areas in the image (eyes, nose, mouth) and temporal anomalies in the audio. Testing on public datasets and simulated data yielded 80.80% accuracy (audio) and 80.75% accuracy (images), with 84.80% precision, 78.50% recall, and 81.50% F1-score at batch size 16. This research contributes to the development of academically transparent deepfake detection methods and supports information security in practice. However, improvements in accuracy and resilience to sophisticated manipulation are still needed.

Keywords: *Deepfake, Deep Learning, Convolutional Neural Networks, Capsule Networks, Explainable AI.*