**ABSTRACT**

With the rise of advanced computer vision and natural language processing techniques, deep learning (DL) models have enabled the creation of highly realistic deepfake content, including manipulated images, videos, and audio. These deepfakes are increasingly used to spread misinformation, propaganda, and disinformation, posing significant threats to personal and organizational reputations. Existing research has predominantly focused on detecting deepfake images and videos, often overlooking the importance of audio-visual synchronization. However, MTCNN (Multi-Task cascaded Convolutional Networks) are now capable of producing highly convincing audio-visual deepfakes, necessitating robust multi-modal detection approaches.

This project presents a comprehensive study of multi-modal deepfake detection using a combination of audio and visual cues to improve detection accuracy. We explore state-of-the-art machine learning (ML) and deep learning models that analyse both speech patterns and facial movements to identify inconsistencies. By leveraging multi-modal fusion techniques, our approach enhances the detection of subtle manipulations that are challenging to detect using single-modality methods. Additionally, we provide a comparative analysis of public datasets suitable for multi-modal deepfake detection, including those containing synchronized audio and video manipulations.

We also discuss implementation challenges such as data imbalance, computational complexity, and the need for real-time detection capabilities. To address these challenges, we explore optimized architectures, including attention mechanisms and transformer-based models, for efficient feature extraction and fusion. A unique case study, IBMM, is presented to demonstrate the effectiveness of our proposed approach.