Introduction

Digital manipulation technology has evolved rapidly with the advancement of artificial intelligence and deep learning. These technologies enable the creation of realistic visual content that is difficult to distinguish from the real thing. One of the most striking results of these developments is the rise of deepfakes, which utilize sophisticated algorithms to produce fake images and videos with a very high degree of realism. One of the most commonly used types of deepfake is face swap, where a person's face is manipulated to resemble another person's face. Face swap can produce videos of a person performing actions or engaging in situations that they never actually did. This technology can be used for a variety of purposes, from entertainment to spreading misinformation. However, its potential misuse for criminal purposes, including the spread of fake news and manipulation of an individual's image, raises serious privacy and security concerns [1].

Detection of deepfakes, especially face swaps, is becoming increasingly important to prevent the negative impact they can have. One of the main challenges in detecting face swaps is that the pattern of manipulation is often very subtle. These manipulations are not always obvious at first glance, as details such as skin color alignment, lighting, and facial texture are often made to appear as if they blend in perfectly with other visual elements. In addition, increasingly sophisticated deepfake technologies, such as those generated by Generative Adversarial Networks (GANs), often leave little to no detectable visual trace [2]. To overcome these challenges, detection methods that can analyze spatial and temporal patterns simultaneously are needed.

Various approaches have been proposed for deepfake detection. In 2021, Tariq et al. [3] developed a Residual Network-based Convolutional LSTM model (CLRNet) that utilizes spatial and temporal information to detect deepfake videos. The model achieved a detection accuracy of 93.86% on the DeepFake-in-the-Wild dataset. However, the model had difficulty detecting deepfake patterns in videos with low resolution or high variations in lighting, which makes it less robust under real-world conditions, where videos with varying qualities and lighting are prevalent. Furthermore, it struggled with generalization to unseen types of manipulations. This limitation highlights the need for more adaptable models. Moreover, CLRNet lacks a mechanism to focus on the most informative time steps across frames, which is essential for detecting subtle manipulation patterns.

In 2022, Saikia et al. proposed a CNN and LSTM-based approach with optical flow features, achieving 91.21% accuracy on the FaceForensics++ dataset. Although effective, this approach has a high processing time due to the complex optical flow analysis, making it unsuitable for real-time detection [4]. The reliance on optical flow features also limits its robustness under different video resolutions and poses, leading to degraded performance

when processing high-motion videos. Additionally, the model does not fully exploit temporal dependencies, which are crucial for detecting dynamic facial manipulations.

In 2019, Rössler et al. proposed a method for detecting deepfakes and face swaps, utilizing the FaceForensics++ dataset. The model developed in this study achieved an accuracy of 90% for detecting face swap videos [5]. This result demonstrates the model's effectiveness in identifying face swaps, a key challenge in deepfake detection, although the model still faces challenges in terms of computational efficiency and robustness to different video qualities and manipulations. While this model outperforms previous methods in terms of accuracy, it still suffers from issues related to computational efficiency, especially for large-scale video datasets. Additionally, like other CNN-based approaches, it may not capture long-range dependencies and temporal features as effectively as models that integrate temporal analysis, which are essential for detecting complex deepfake manipulations across multiple frames.

Recent work by Coccomini et al. (2023) proposed combining EfficientNet B0 and Vision Transformers (ViT) to detect manipulations in video. Their method, integrating Efficient ViT and Convolutional Cross ViT, achieved promising results with 80% accuracy on FaceForensics++ sub-datasets, outperforming traditional ViT models. This highlights the potential of ViT, especially when combined with convolutional techniques, for detecting deepfake and face swap [6]. However, the weak supervision approach may limit accuracy with partially labeled data and might not fully exploit the rich temporal information in video data.

To address the challenges of detecting subtle manipula tions in face swap deepfake videos, this research proposes a combined approach by integrating Swin Transformer and BiLSTM with an attention mechanism. The Swin Transformer, as an innovation of the Vision Transformer (ViT), excels in capturing both global and local spatial features with high com putational efficiency through its window-based self-attention mechanism [9]. Compared to CNN, the Swin Transformer is better suited for handling high-resolution images and extract ing more complex visual patterns. Meanwhile, the BiLSTM with an attention mechanism effectively captures temporal patterns between frames, enabling deeper analysis of dynamic relationships in face swap videos.

Based on the application of the Swin Transformer and BiLSTM in multi-task recognition for gesture and identity in FMCW radar applications, as demonstrated in SwinFMCW [17], this research extends the potential of the architecture to the deepfake detection domain. By integrating spatial and temporal analysis through the advantages of the Swin Transformer, BiLSTM, and attention mechanisms, this research aims to improve the accuracy of deepfake detection, especially face swap, on datasets with complex variations in face pose, lighting, and motion. This demonstrates the versatility of this deep learning approach across various application domains.