A Systematic Literature Review of Deepfake Face Image Detection with Transfer Learning Techniques

Waruni Wimalasena
Software Engineering Teaching Unit
Faculty of Science
University of Kelaniya
Kelaniya, Sri Lanka
rawlw241@kln.ac.lk

Hansika Herath
Software Engineering Teaching Unit
Faculty of Science
University of Kelaniya
Kelaniya, Sri Lanka
hmkhh241@kln.ac.lk

Isuru Hewapathirana
Software Engineering Teaching Unit
Faculty of Science
University of Kelaniya
Kelaniya, Sri Lanka
ihewapathirana@kln.ac.lk

***Abstract* — Artificial intelligence-based fake content generation, also known as deepfake media content, is spreading rapidly with the advancement of technology. This can manipulate or completely change the real media by threatening the authenticity of the data. With the availability of major amounts of data, deepfake face image generation has become an emerging issue in today's digitalized world. Therefore, it is crucial to understand the most effective methods to detect this deepfake face image content. This paper provides a comprehensive discussion about how transfer learning can be used in this specific area to detect deepfake face images through a systematic literature review of (49) papers using the PRISMA method. The paper addresses the gap in up-to-date detailed analysis of transfer learning methods in this area which emphasizes the effectiveness of the process of deepfake image detection by studying papers in the past decade. This paper presents a comprehensive discussion of the models that have been using this area along with the transfer learning approaches that have been used. The key findings highlight that the most common transfer learning approach for deepfake image detection is CNN combined and the most effective model for deepfake image detection in transfer learning is Fine-tuned models. This paper also highlights the challenges and possible improvements in the area of deepfake face image detection with transfer learning techniques.**

***Keywords* —Deepfake image detection, Face recognition, Systematic literature review, Transfer learning**

I. Introduction

Rapid advancements in artificial intelligence (AI), Particularly generative AI, have significantly influenced every aspects of human life including digital media [1] .This impact is intense in fields where large datasets are readily available. With the increasing accessibility and sophistication of AI Technologies, AI generated digital media is experiencing exponential growth in society [2]. These synthetic-generated media have become a threat to society in many aspects. Especially, due to the sensitive nature of the data, deepfake face image generation has emerged as a growing concern [3]. This challenges the authenticity of digital media and threats to privacy, security, and trust in online platforms. Addressing these issues requires the development of effective detection methods to counter the misuse of deepfake technology.

Face image manipulation using AI technologies can be categorized into four primary types: (i) entire face synthesis, (ii) identity swap (DeepFakes), (iii) attribute manipulation, and (iv) expression swap [4]. There are many approaches to detect deepfake image generation, which can be broadly classified into three categories: (i) deep learning-based techniques, (ii) hybrid techniques such as combining deep learning with machine learning or statics-related techniques, and (iii) alternative techniques such as blockchain-related approaches. However, research indicates that among these categories, that the most prominent method is deep learning based techniques. To enhance the accuracy of these models, researchers have increasingly focused on transfer learning techniques [5], which have given promising results including improved overall process efficiency. For example, transfer learning techniques have shown greater results even with limited datasets.

Previous research has been using several models for transfer learning approaches such as ImageNet, MobileNet, and RestNet architectures [6]. While each model has demonstrated promising results, suggesting significance potential for transfer learning in deepfake image detection [7], a comprehensive analysis of their relative effectiveness remains necessary. Although numerous literature reviews have addressed overall deepfake detection, deepfake images are the most common issue regarding deepfake media generation. There exists a gap in a systematic literature review specifically of deepfake detection with this promising approach for transfer learning techniques [8].

This paper addresses this research gap by conducting a systematic literature review of research papers from the past decade. The paper has used the PRISMA method to ensure a rigorous and transparent review process. The research aims to address four fundamental questions regarding transfer learning within the domain of deepfake face image detection, emphasizing its significant impact on deep fake efficacy.

979-8-3315-2326-8/25/$31.00