IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 25, 2022, accepted February 16, 2022, date of publication February 24, 2022, date of current version March 10, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3154404

Deepfake Detection: A Systematic Literature Review

MD SHOHEL RANA[1,2], (Member, IEEE), MOHAMMAD NUR NOBI[3], (Member, IEEE), BEDDHU MURALI[2], AND ANDREW H. SUNG[2], (Member, IEEE)

[1]Department of Computer Science, Northern Kentucky University, Highland Heights, KY 41099, USA
[2]School of Computing Sciences and Computer Engineering, The University of Southern Mississippi, Hattiesburg, MS 39401, USA
[3]Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX 78249, USA

Corresponding author: Md Shohel Rana (ranam2@nku.edu)

This work was supported in part by Northern Kentucky University and the University of Southern Mississippi.

ABSTRACT Over the last few decades, rapid progress in AI, machine learning, and deep learning has resulted in new techniques and various tools for manipulating multimedia. Though the technology has been mostly used in legitimate applications such as for entertainment and education, etc., malicious users have also exploited them for unlawful or nefarious purposes. For example, high-quality and realistic fake videos, images, or audios have been created to spread misinformation and propaganda, foment political discord and hate, or even harass and blackmail people. The manipulated, high-quality and realistic videos have become known recently as Deepfake. Various approaches have since been described in the literature to deal with the problems raised by Deepfake. To provide an updated overview of the research works in Deepfake detection, we conduct a systematic literature review (SLR) in this paper, summarizing 112 relevant articles from 2018 to 2020 that presented a variety of methodologies. We analyze them by grouping them into four different categories: deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques. We also evaluate the performance of the detection capability of the various methods with respect to different datasets and conclude that the deep learning-based methods outperform other methods in Deepfake detection.

INDEX TERMS Deepfake detection, video or image manipulation, digital media forensics, systematic literature review.

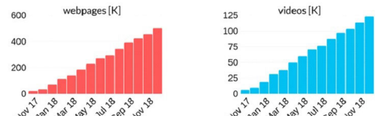
I. INTRODUCTION

The notable advances in artificial neural network (ANN) based technologies play an essential role in tampering with multimedia content. For example, AI-enabled software tools like FaceApp [1], and FakeApp [2] have been used for realistic-looking face swapping in images and videos. This swapping mechanism allows anyone to alter the front look, hairstyle, gender, age, and other personal attributes. The propagation of these fake videos causes many anxieties and has become famous under the hood, Deepfake.

The term ''Deepfake'' is derived from ''Deep Learning (DL)'' and ''Fake,'' and it describes specific photo-realistic video or image contents created with DL's support. This word was named after an anonymous Reddit user in late 2017, who applied deep learning methods for replacing a person's face in pornographic videos using another person's face and created photo-realistic fake videos. To generate such counterfeit videos, two neural networks: (i) a generative network and (ii) a discriminative network with a FaceSwap technique were used [3], [4]. The generative network creates fake images using an encoder and a decoder. The discriminative network defines the authenticity of the newly generated images. The combination of these two networks is called Generative Adversarial Networks (GANs), proposed by Ian Goodfellow [5].

Based on a yearly report [6] in Deepfake, DL researchers made several related breakthroughs in generative modeling. For example, computer vision researchers proposed a method known as Face2Face [7] for facial re-enactment. This method transfers facial expressions from one person to a real digital 'avatar' in real-time. In 2017, researchers from UC Berkeley presented CycleGAN [8] to transform images and videos into different styles. Another group of

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.



on surveying selected literature focusing on either detection methods or performance analysis. However, a more comprehensive overview of this research area will be beneficial in serving the community of researchers and practitioners by providing summarized information about Deepfake in all aspects, including available datasets, which are noticeably