Analysis of Deepfake Detection Models Against Deepfake Crimes

Sechan Lee[1], Sooyon Seo[2], Jaeho Hwang[2], Minseok Hur[2], Dongmin Kim[3], Aram Kim[4] and Moohong Min[5]

[1] Department of Computer Education, Sungkyunkwan University, Seoul, Republic of Korea
chan1031@g.skku.edu
[2] Department of Immersive Media Engineering, Sungkyunkwan University, Seoul, Republic of Korea
{sooyon1119, ashiereki, alexhur3535}@skku.edu
[3] Department of Data Science, Sungkyunkwan University, Seoul, Republic of Korea
ehdals5744@g.skku.edu
[4] Department of Information Security, University of Suwon, Suwon, Republic of Korea
aramkim@suwon.ac.kr
[5] Department of Computer Education/Social Innovation Convergence Program, Sungkyunkwan University, Seoul, Republic of Korea
iceo@skku.edu

Abstract

This paper analyzes and compares detection models designed to combat various crimes arising from the advancement of deepfake technology. Deepfake technology enables the realistic synthesis of facial features, voices, and actions to create fake videos, which can lead to social chaos such as the spread of fake news, financial fraud and sexual crimes. To counter these threats, this paper introduces key deepfake detection models, including XceptionNet, EfficientNet-B7, and the CNN-LSTM hybrid model, and compares their specialization and drawbacks. Furthermore, the importance of deepfake detection technology is emphasized, and the need for legal regulations and public awareness to prevent deepfake-related crimes is proposed.

1 Introduction

In recent years, artificial intelligence technology has made remarkable advancements, significantly contributing to the convenience of our daily lives. However, as this technology has developed and become accessible to the public, cases of its misuse have increased. Among these, deepfake technology, which can precisely synthesize videos or audio, has been exploited for various criminal

activities. In particular, a deepfake-related sexual crime that occurred in 2024 in South Korea through "Telegram" caused significant social uproar. Additionally, deepfake technology, if misused, can lead to substantial social chaos through the spread of misinformation, fake news, and financial fraud involving identity theft. These cases demonstrate that the misuse of deepfake technology poses a severe threat to society. This paper aims to examine the current state of deepfake detection technologies in response to such misuse and propose strategies for preventing deepfake-related crimes.

2 Theoretical Background

2.1 Deepfake

Deepfake refers to a technology that utilizes artificial intelligence and deep learning to realistically synthesize a person's face, voice, and actions. It primarily employs GAN (Generative Adversarial Network) [1]. GAN is a type of unsupervised learning technique where a generator and a discriminator compete against each other to learn from the data.

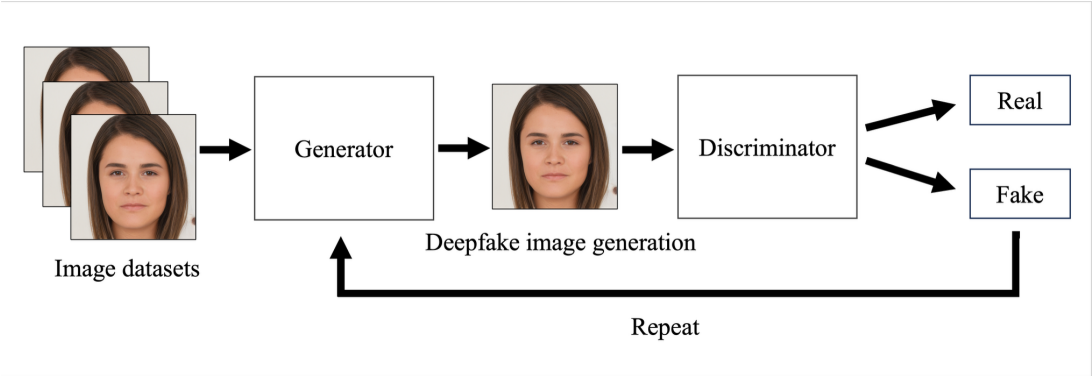

Figure 1: The process of generating deepfake images using GAN

The process of generating deepfake images is as follows. First, a dataset of facial images of the person to be used for the deepfake is required. Then, the GAN training begins. The generator creates deepfake images, while the discriminator attempts to distinguish whether the images generated by the generator are real or fake. If the discriminator identifies an image as fake, the generator improves it and creates a more refined fake image. This process is repeated, resulting in increasingly realistic deepfake composite images.

2.2 Misuse Cases of Deepfake Technology

Deepfake crimes refer to offenses that involve the misuse of deepfake technology. This includes creating fake news, identity theft, and sexual crimes, all of which can lead to significant social disruption. Notable examples of such misuse include the following:

During the 2022 Ukraine-Russia war, a deepfake video of Ukrainian President Volodymyr Zelensky surrendering to Russia was circulated on Twitter, causing social unrest [2]. Additionally, there have been cases where deepfake videos were created to impersonate trusted individuals, leading to financial fraud that the victims did not suspect [3]. Moreover, in 2024, South Korea was rocked by a significant scandal involving the distribution of deepfake pornography through Telegram. Unlike previous deepfake crimes that typically targeted celebrities, this case was particularly shocking as it involved ordinary middle and high school students, as well as teachers. The perpetrators collected

selfies from the victims' social media accounts, used deepfake technology to manipulate the images, and then distributed the resulting content in group chats on Telegram. Over 200 victims were identified, and the criminal activities had been organized since 2020. The widespread victimization in this case highlights the fact that, as deepfake technology advances, it becomes accessible even to those with limited knowledge of artificial intelligence, making it easier for anyone to misuse. In fact, deepfake software like "Faceswap" and applications like "Reface" are readily available, allowing anyone to easily create deepfake content.

3 Deepfake Detection Model

The current state of deepfake technology has advanced to such a degree that it is difficult to distinguish deepfakes with the naked eye. Therefore, to determine whether the content is a deepfake, it is necessary to employ detection techniques that also utilize artificial intelligence through deepfake detection models.

3.1 Deepfake Detection Methods

The process of distinguishing deepfakes using a deepfake detection model is illustrated in Figure 2 below.

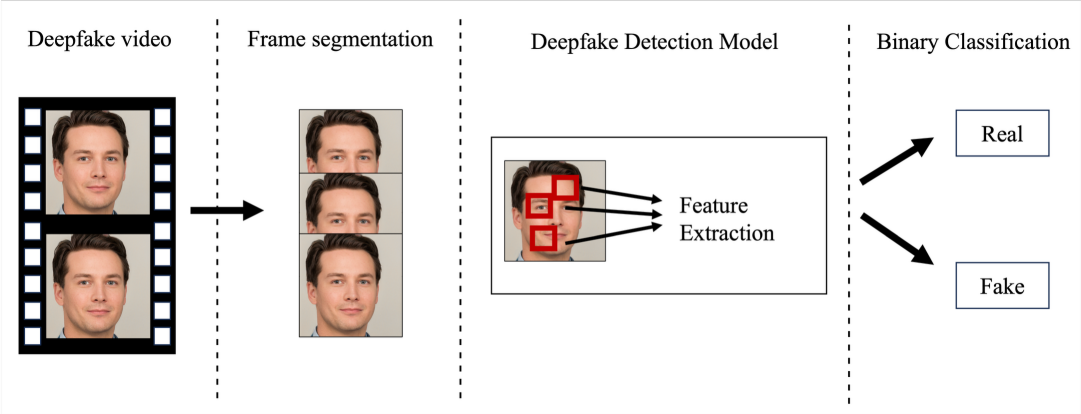

Figure 2: The process of detecting deepfakes

The video is divided into individual frames, and each frame is analyzed using a deepfake detection model to identify characteristics unique to fake videos. Multiple characteristics of deepfake videos are used.

First, Eye Aspect Ratio (EAR): One of the most common characteristics of deepfake videos is the unnatural blinking of the eyes. Since deepfakes often use static images or images where the eyes are open, the natural frequency of human blinking does not align with that in deepfake videos, where the blinking rate is typically faster. The blinking frequency. can be detected using the EAR (Eye Aspect Ratio) algorithm [4]. This algorithm calculates the vertical and horizontal distances of eye landmarks and determines whether the eyes are open or closed based on these calculations. The formula for the EAR algorithm is as follows:

$$EAR = \frac{|P_2 - P_6| + |P_3 - P_5|}{2 \times |P_1 - P_4|} \quad (1)$$

P_1 to P_6 represent specific landmark points around the eye. P_1 and P_4 correspond to the coordinates on the horizontal axis of the eye (the outer corners), P_2 and P_6 correspond to the vertical axis coordinates (the upper and lower eyelids), and P_3 and P_5 represent the vertical axis coordinates at the center of the eye. When the eyes are open, the EAR value is relatively high, and when the eyes are closed, the EAR value decreases, allowing the blinking frequency to be determined.

Second, **Visual Artifact Detection**: This method identifies subtle differences and inconsistencies between the face and the surrounding background during the face synthesis process in deepfake generation. It distinguishes deepfakes by detecting inconsistencies in resolution, pixel anomalies, color, and brightness.

Third, **Audio-Video Mismatch**: In some deepfake videos, the movements of a person's lips do not align with the video. Lip-sync analysis is used to detect discrepancies between the audio and video, helping to identify deepfakes.

These characteristics are then combined, and a binary classification is performed to determine whether the content is a deepfake. The accuracy and training time of deepfake detection vary depending on the model used. The most commonly used deepfake detection models include the following.

3.2 XceptionNet

XceptionNet [6], proposed by Chollet in 2017, is a model that stands for "Extreme Inception" and is an extended version of the original Inception architecture. This model demonstrated highly effective detection performance, placing in the top three of the 2019 Kaggle Deepfake Detection Challenge [5]. XceptionNet employs "Depthwise separable convolutions" to reduce model complexity while enhancing performance. Depthwise separable convolution, the core component of XceptionNet, differs from traditional convolution by separating the process into two parts: "Depthwise Convolution" and the "Pointwise Convolution". The method for detecting deepfakes is illustrated in Figure 3.

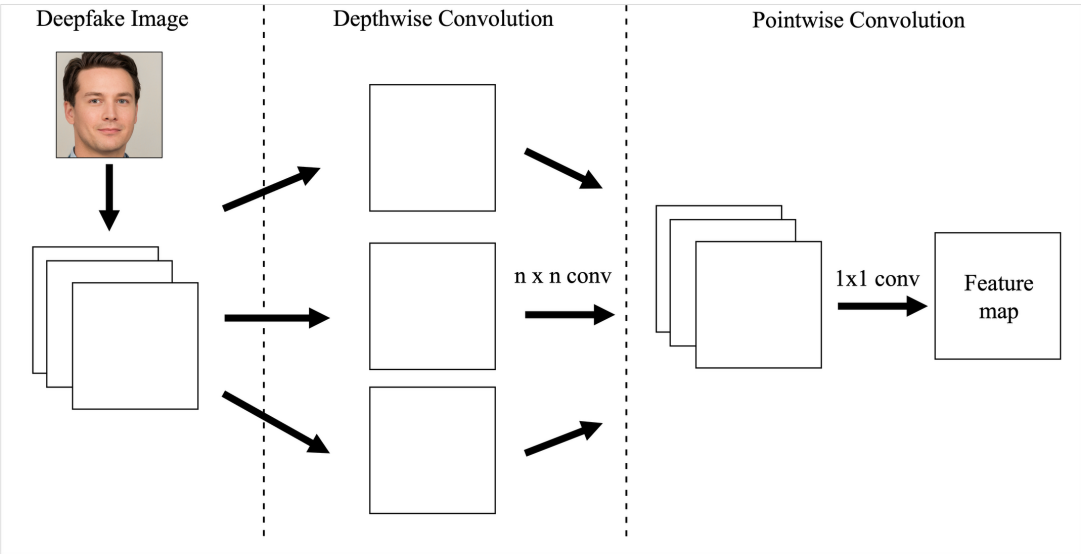

Figure 3: XceptionNet's Depthwise Seperable Convolution

First, Depthwise Convolution performs convolution operations independently on each channel, extracting spatial features without interaction between channels. Then, in the Pointwise Convolution,

a 1x1 convolution is applied to combine the outputs of each channel to generate a new feature map. Through this structure, XceptionNet reduces model complexity while enabling efficient feature extraction. In deepfake detection, XceptionNet can extract subtle visual effects or abnormal patterns through depthwise separable convolutions. It can learn complex patterns with fewer parameters, providing consistent detection performance across various deepfake techniques. Compared to traditional CNN(Convolution Neural Network) models like VGGNet or ResNet [6], it offers better performance.

3.3 EfficientNet-B7

As its name suggests, the EfficientNet model [7] is designed to maximize efficiency and performance. EfficientNet achieves this by using a compound scaling technique to simultaneously adjust the network's width, depth, and input image resolution, thereby maximizing efficiency.
The architecture of EfficientNet Model is Figure 4.

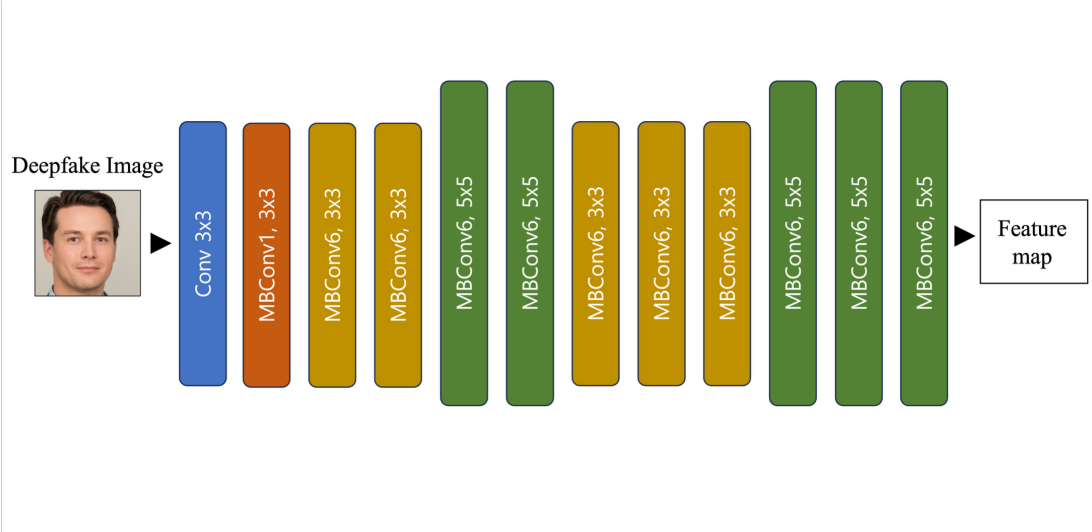

Figure 4: Architecture of EfficientNet

Previous models were unable to adjust network width, depth, and input image resolution all at once to improve model performance. However, EfficientNet balances and adjusts these three factors together through compound scaling, allowing the model's performance to be maximized beyond previous levels. To perform compound scaling, the three elements need to be expanded simultaneously and in an optimized manner, which is done using the following formula. (2)

$$depth = \alpha^{\phi}, width = \beta^{\phi}, resolution = \gamma^{\phi} \quad (2)$$

This formula determines the extent of network scaling simultaneously through expansion coefficients for each depth, width, and resolution constant. EfficientNet-B7 is one of the largest EfficientNet models, significantly adjusting the network's width, depth, and image resolution to achieve very high performance. There are several advantages of using EfficientNet-B7 for deepfake detection. Firstly, it allows for high-resolution image processing. EfficientNet-B7 can process images at a resolution of 600x600, enabling the extraction of subtle features that appear in high-resolution deepfake images. Additionally, its high efficiency provides excellent performance with fewer parameters and lower computational load, making it highly suitable for large-scale data processing

and real-time detection. Moreover, with compound scaling, the model's size and performance can be adjusted according to the context, allowing flexible application across various datasets.

3.4 CNN-LSTM Hybrid Model

The combined model of CNN-LSTM hybrid model is also effective in deepfake detection [8]. By merging CNN's powerful spatial feature extraction capabilities with LSTM(Long Short-Term Memory)'s ability to process temporal data, this model can be useful for handling complex tasks like deepfake detection. The process of how the CNN-LSTM hybrid model detects deepfakes is illustrated in Figure 4.

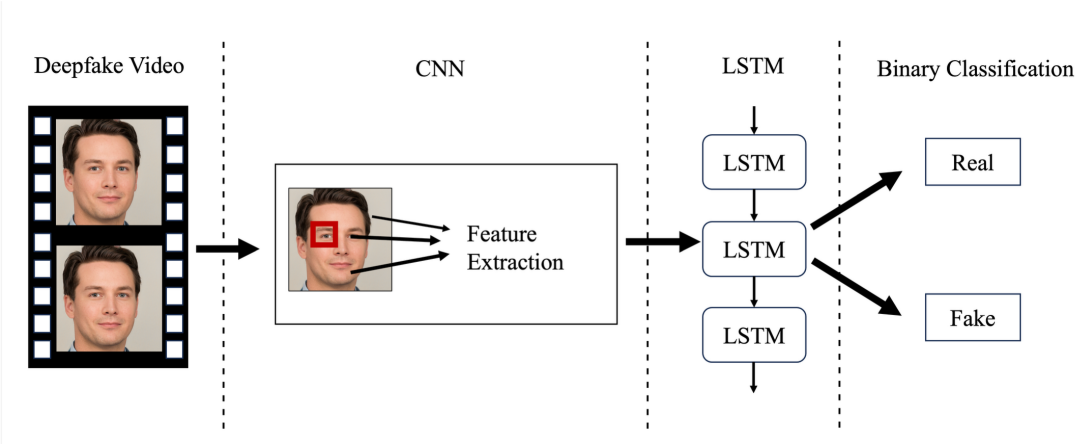

Figure 5: CNN-LSTM hybrid model architecture

First, the video is split into individual frames. Each frame is analyzed by the CNN to extract spatial features of the face, such as the eyes, nose, and mouth. After passing through the CNN, each frame is converted into a fixed-size vector, which is then passed to the LSTM. Since the LSTM can learn from sequential data, it captures the temporal dependencies among the received frames to identify the unnaturalness of the deepfake video. For example, it can detect unnatural blinking, mismatches between mouth movements and speech, and other inconsistencies. CNN-LSTM hybrid model is effective for detecting deepfake videos. Because Deepfakes inevitably exhibit unnaturalness in consecutive frames, making it possible to efficiently assess these unnatural aspects through the CNN-LSTM hybrid model approach.

4 Comparison of Deepfake Detection Methods

The three deepfake detection models described earlier are widely used and demonstrate high accuracy, but each model has its own strengths and weaknesses, as shown in Table 1.
The analysis results of the following models can help in choosing the most suitable model when developing a deepfake detection system. For instance, in the case of deepfake video analysis, a model combining CNN and LSTM can be created, while EfficientNet could be considered as a candidate for cases involving smaller datasets. Selecting an appropriate model tailored to the specific situation is the best way to enhance the efficiency of deepfake detection.

Table 1: Comparison of deepfake detection models

Model	Specialization	Drawbacks
XceptionNet	Detects fine visual artifacts effectively	High memory usage, long training time
EfficientNet-B7	Excels at high-resolution image detection	Complex parameter tuning required
CNN-LSTM hybrid model	Effective in detecting temporal inconsistencies	Slower processing speed

5 Conclusion

Deepfake generation technology has advanced to the point where it is virtually indistinguishable to the naked eye. Moreover, services based on websites or applications are now easily accessible to the public. Therefore, continuous research and development of detection models are essential to detecting numerous illegal deepfake creations. Fortunately, even as deepfake technology evolves, accumulating relevant data will allow the training of detection models, thereby enhancing their performance. This ongoing battle between deepfakes and detection technologies resembles a never-ending struggle, akin to a fight between a sword and a shield. In addition to developing deepfake detection technology, it is crucial to strengthen relevant legislation and impose sanctions on services facilitating their creation [1]. Increasing penalties for users who cause social disruption through deepfakes is vital for changing public perception. Additionally, preventing services from operating in cases that can lead to social chaos—such as sexual crimes or fake news—will be a key strategy in preventing deepfake-related offenses.

6 Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea goverment(MSIT) (No.RS-2024-00439139, Development of a Cyber Crisis Response and Resilience Test Evaluation Systems). And this work was supported by the MSIT(Ministry of Science and ICT), Korea, under the Graduate School of Metaverse Convergence support program(IITP-2024-RS-2023-00254129) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)

References

[1] Westerlund, M. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11): 39–52, 2019.
[2] Wakefield Jane. *Deepfake Presidents Used in Russia-Ukraine War*. BBC. https://www.bbc.com/news/technology-60780142, 2022.
[3] Sjouwerman Stu. *Deepfake Phishing: The Dangerous New Face Of Cybercrime*.

Forbes. https://www.forbes.com/councils/forbestechcouncil/2024/01/23/deepfake-phishing-the dangerous-new-face-of-cybercrime/, 2024.
[4] T. Jung, S. Kim and K. Kim, DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern, IEEE Access, 8: 83144-83154, 2020.
[5] V. L. L. Thing. Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers. 2023 IEEE International Conference on Cyber Security and Resilience (CSR), pages 246-253, 2023.
[6] P. Joshi and N. V. Deep Fake Image Detection using Xception Architecture. 2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), pages 533-537, 2024.
[7] A. A. Pokroy and A. D. Egorov, EfficientNets for DeepFake Detection: Comparison of Pretrained Models. 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), pages 598-600, 2021.
[8] P. Saikia, D. Dholaria, P. Yadav, V. Patel and M. Roy. A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features. 2022 International Joint Conference on Neural Networks (IJCNN), pages 1-7, 2022.