Deepfake Detection: A Tutorial

Md Shohel Rana[*]
Stephen F. Austin State University
Nacogdoches, Texas, USA
md.rana@sfasu.edu

Andrew H. Sung
The University of Southern Mississippi
Hattiesburg, Mississippi, USA
andrew.sung@usm.edu

ABSTRACT

This tutorial presents developments on the detection of Deepfakes, which are realistic images, audios and videos created using deep learning techniques. Deepfakes can be readily used for malicious purposes and pose a serious threat to privacy and security. The tutorial summarizes recent Deepfake detection techniques and evaluates their effectiveness with respect to several benchmark datasets. Our study finds that no single method can reliably detect all Deepfakes and, therefore, combining multiple methods is often necessary to achieve high detection rates. The study also suggests that more extensive and diverse datasets are needed to improve the accuracy of detection algorithms. A taxonomy of Deepfake detection techniques is introduced to aid future research and development in the field. We conclude by calling for the development of more effective Deepfake detection methods and countermeasures to combat this evolving and spreading threat.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**.

KEYWORDS

Deepfakes, Deep learning, Taxonomy, Security and Privacy

ACM Reference Format:
Md Shohel Rana and Andrew H. Sung. 2023. Deepfake Detection: A Tutorial. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics (IWSPA '23), April 26, 2023, Charlotte, NC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3579987.3586562

1 INTRODUCTION

The rapid advancement of artificial neural network (ANN) based technologies has made it easier than ever to manipulate multimedia content. AI-enabled software tools like FaceApp [2] and FakeApp [3] have been used for realistic-looking face swapping in videos, leading to a surge of concern and anxiety over the propagation of fake videos. This phenomenon, known as Deepfake, refers to specific photo-realistic video or image content created with deep learning (DL) techniques. Deepfakes have numerous malicious uses, including spreading misinformation, creating political discord, or perpetrating various cybercrimes.

In spite of the growing concern over the spread of Deepfakes, DL researchers have continued to make significant advances in generative modeling. For example, computer vision researchers have proposed methods for facial re-enactment [9], image and video transformation into different styles [10], and lip movement synchronization [8]. Deepfakes have been widely used in pornography early on, with multiple platforms posting thousands of Deepfake videos; more recently Deepfakes have been exploited for financial fraud and extortion [1]. Recognizing the emerging threat of Deepfakes, the field of Deepfake detection has attracted considerable attention from academic researchers, financial services, cybersecurity practitioners, and law enforcement alike, and hundreds of papers have been published in various venues in just recent years.

In a recent IEEE Access paper [6], we presented a systematic literature review on Deepfake detection research regarding current tools, techniques, and datasets. We classify these studies into four categories: deep learning-based techniques, classical machine learning-based methods, statistical techniques, and blockchain-based techniques. To evaluate the performance of these techniques, we use several benchmark datasets, including FaceForensics++ (FF++) [7], Deepfake Detection Challenge (DFDC) [5], and Celeb-DF [4]. We compare the detection capability of different techniques on these datasets using standard evaluation metrics such as precision, recall, and F1-score. The best-performing techniques reported in the study achieved a detection accuracy of over 99% on the FF++ and over 93% on the DFDC.

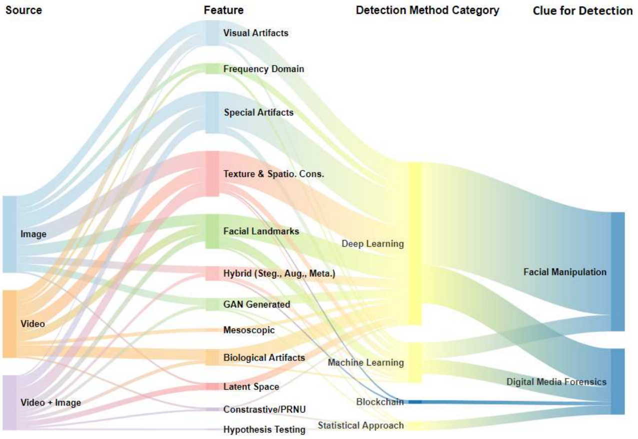

Figure 1: Taxonomy of Deepfake detection techniques. This taxonomy classifies the detection algorithms according to the media, the features used, the detection methods, and the clue for the detection.

IWSPA '23, April 26, 2023, Charlotte, NC, USA

ACM ISBN 979-8-4007-0099-6/23/04…$15.00
https://doi.org/10.1145/3579987.3586562

We also introduced a taxonomy (see Figure 1) that classifies Deepfake detection techniques into four categories based on their features. Our taxonomy provides an overview of the different categories and their related features, which will be useful for future research and development of Deepfake detection methods and aims to describe and analyze common grounds and the diversity of approaches in current practices.

Our taxonomy is based on Detection (i.e., Digital Media Forensics, and Face Manipulation), Methods indicate the algorithmic category (ML: Machine Learning, DL: Deep Learning, STAT: Statistical method), Models represent types of model (see Figure 2) where (DL: (CNN: Convolutional Neural Network, RNN: Recurrent Neural Network, RCNN: Regional Convolutional Neural Network), ML: (SVM: Support Vector Machine, RF: Random Forest, MLP: Multilayer Perceptron Neural Network, LR: Logistic Regression, k-MN: K means clustering, XGB: XGBoost, ADB: AdaBoost, DT: Decision Tree, NB: Naive Bayes, KNN: K-Nearest Neighbour, DA: Discriminant Analysis), STAT: (EM: Expectation Maximization, CRA: Co-relation Analysis)), Features (Special Artifacts, Visual Artifacts, Biological Artifacts, Face Landmarks, Spatio-temporal Consistency, Texture, Frequency Domain Analysis, Latent Feature, Generative Adversarial Network based feature, Mesoscopic features, Intra-frame inconsistency, Constrastive and Photoresponsive PRNU pattern, Image Metadata, Augmentation & Steganalysis, Others), Datasets (FaceForensics++ (FF++), Deepfake Detection: DeepFake Forensics V1 (CELEB-A), DeepFake Forensics V2 (CELEB-DF), Deepfake Detection Challenge (DFDC), Deepfake-TIMIT, DeeperForensics-1.0 (DF-1.0), Deep Fakes (DFS), Fake Faces in the Wild (FFD), FakeET, Face Shifter, Deepfake (DF), Inconsistent Head Poses (UADFV), Tampered Face (MANFA), and others).

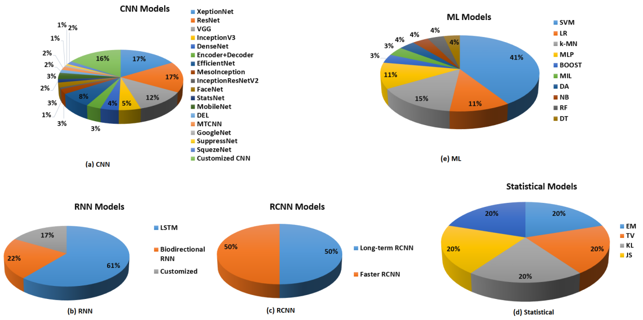

Figure 2: The allocation of categories of detection models. (a) CNN (b) RNN (c) RCN (d) Statistical, and (e) ML.

2 CONCLUSION

This tutorial provides a comprehensive overview of the current state of Deepfake detection research, including available datasets and performance analysis, with the literature covered in the study mostly focused on Deepfake videos.

Our study reveals that there is no single method that can reliably detect all Deepfakes, and that combining multiple methods would seem essential to achieving high detection accuracy. Additionally, we found that current datasets for training and evaluating Deepfake detection methods are mostly inadequate due to their limited size and lack of diversity, so more extensive and diverse datasets are needed for training, testing and validation in building new models to improve the accuracy of detection.

Multimedia forensics concerns the analysis of manipulated multimedia contents and has been an important subject of digital forensics; currently, detection methods and countermeasures for Deepfakes are active research topics of multimedia forensics. The recently introduced generative AI tools (for answering questions, composing essays, etc.) have brought forth interesting challenges of how to detect this kind of AI-generated material. We hope that the tutorial will be useful for researchers and practitioners of multimedia forensics and motivate the development of more effective methods to combat the spreading threats to security and privacy posed by manipulated and/or AI-synthesized multimedia.

REFERENCES

[1] Jon Bateman. 2020. Fakeapp. Webpage. (Aug. 2020). Retrieved March 11, 2023 from https://carnegieendowment.org/2020/08/10/get-ready-for-deepfakes-to-be-used-in-financial-scams-pub-82469.

[2] 2017. Faceapp. Application. (2017). Retrieved March 11, 2023 from https://www.faceapp.com/.

[3] 2018. Fakeapp. Application. (2018). Retrieved March 11, 2023 from https://www.malavida.com/en/soft/fakeapp/.

[4] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: a large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3204–3213. DOI: 10.1109/CVPR42600.2020.00327.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 3730–3738. DOI: 10.1109/ICCV.2015.425.

[6] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H. Sung. 2022. Deepfake detection: a systematic literature review. *IEEE Access*, 10, 25494–25513. DOI: 10.1109/ACCESS.2022.3154404.

[7] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. Faceforensics++: learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–11. DOI: 10.1109/ICCV.2019.00009.

[8] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36, 4, Article 95, 13 pages. DOI: 10.1145/3072959.3073640.

[9] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. 2018. Face2face: real-time face capture and reenactment of rgb videos. *Commun. ACM*, 62, 1, 96–104. DOI: 10.1145/3292039.

[10] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251. DOI: 10.1109/ICCV.2017.244.