DEEPFAKE DETECTION A SYSTEMATIC LITERATURE REVIEW

[1]B.HARITHA LAKSHMI, [2]VAIDYA RISHIKA, [3]SANKEPALLI SRI CHANDANA, [4]NALLAMOTHU BHAVYA SREE

[1]Assistant Professor, Department of Information Technology, **MALLA REDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

[2, 3,4] Student, Department of Information Technology, **MALLA REDDY ENGINEERING COLLEGE FOR WOMEN**, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

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.

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 Face App [1], and Fake App [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, Deep fake. The term ``Deep fake" 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 Redd it 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 Face Swap 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 Good fellow [5].

Based on a yearly report [6] in Deep fake, 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 Cycle GAN [8] to transform images and videos into different styles. Another group of scholars from the University of Washington proposed a method to synchronize the lip movement in video with a speech from another source [9]. Finally, in November 2017, the term ``Deep fake'' emerged for sharing porn videos, in which celebrities' faces were swapped with the original ones. In January 2018, a Deepfake creation service was launched by various websites based on some private sponsors. After a month, several websites, including Gfycat [10], Pornhub, and Twitter, banned these services. However, considering the threats and potential risks in privacy vulnerabilities, the study of Deep fake emerged super fast. Rossler *et al.* introduced a vast video dataset to train the media forensic and Deep fake detection tools called Face Forensic [11] in March 2018. After a month, researchers at Stanford University published a method, ``Deep video portraits'' [12] that enables photo-realistic re-animation of portrait videos. UC Berkeley researchers developed another approach [13] for transferring a person's body movements to another person in the video. NVIDIA introduced a style-based generator architecture for GANs [14] for synthetic image generation. According to [6] report, Google search engine could find multiple web pages that contain Deep fake related videos (see Figure 1).We found the following additional information from this report [6]:

_ The top 10 pornographic platforms posted 1,790C Deep fake videos, without concerning pornhub.com, which has removed 'Deep fakes' searches.

_ Adult pages post 6,174 Deep fake videos with fake video content.

_ 3 New platforms were devoted to distributing Deep fake pornography.

_ In 2018, 902 articles were published in arXiv, including the keyword GAN either in titles or abstracts.

_ 25 Papers published on this subject, including non-peer reviews, are investigated, and DARPA funded 12 of them.

Apart from Deep fake pornography, there are many other malicious or illegal uses of Deep fake, such as spreading misinformation, creating political instability, or various cybercrimes. To address such threats, the field of Deep fake detection has attracted considerable attention from academics and experts during the last few years, resulting in many Deep fake detection techniques. There are also some efforts 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 Deep fake in all aspects, including available datasets, which are noticeably missing in previous surveys. Toward that end, we present a systematic literature review (SLR) on Deep fake detection in this paper. We aim to describe and analyze common grounds and the diversity of approaches in current practices on Deep fake detection.

Our contributions are summarized as follows.

_ We perform a comprehensive survey on existing literature in the Deep fake domain. We report current tools, techniques, and datasets for Deep fake detection-related research by posing some research questions.

_ We introduce a taxonomy that classifies Deep fake detection techniques in four categories with an overview of different categories and related features, which is novel and the first of its kind.

_ We conduct an in-depth analysis of the primary studies' experimental evidence. Also, we evaluate the performance of various Deep fake detection methods using different measurement metrics.

_ We highlight a few observations and deliver some guidelines on Deep fake detection that might help future research and practices in this spectrum.

The remainder of the paper is organized as follows: Section II presents the review procedure by defining interest research questions. In Section III, we thoroughly discuss the findings from different studies. Section IV summarizes the overall observations of the study, and we present the challenges and limitations in Section V. Finally, Section VI concludes the paper.

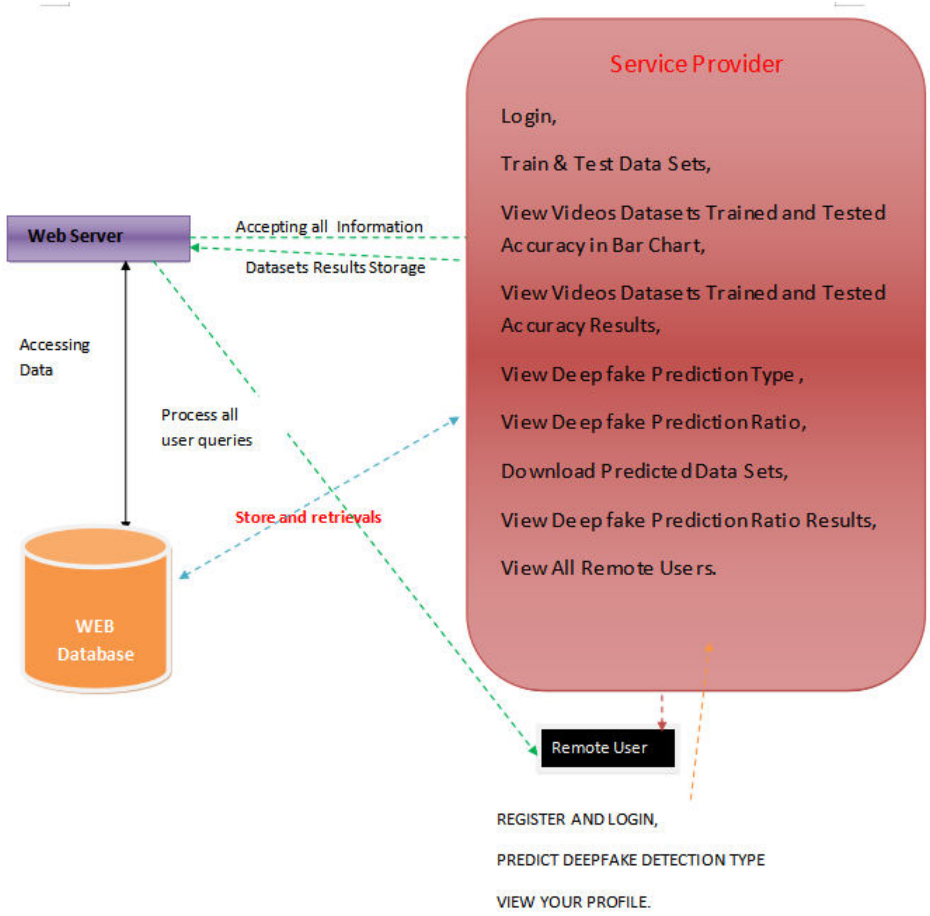
Service Provider
Login,
Train & Test Data Sets,
View Videos Datasets Trained and Tested Accuracy in Bar Chart,
View Videos Datasets Trained and Tested Accuracy Results,
View Deepfake Prediction Type ,
View Deepfake Prediction Ratio,
Download Predicted Data Sets,
View Deepfake Prediction Ratio Results,
View All Remote Users.
Web Server
Accepting all Information
Datasets Results Storage
Accessing Data
Process all user queries
Store and retrievals
WEB Database
Remote User
REGISTER AND LOGIN,
PREDICT DEEPFAKE DETECTION TYPE
VIEW YOUR PROFILE.

Fig1 : System Architecture

II. EXISTING SYSTEM

In [22], the consistency of the biological signs are measured along with the spatial and temporal [23]_[25] directions to use various landmark [26] points of the face (e.g., eyes, nose, mouth, etc.) as unique features for authenticating the legitimacy of GANs generated videos or images. Similar characteristics are also visible in Deepfake videos, which can be discovered by approximating the 3D head pose [27]. In most cases, facial expressions are associated initially with the head's movements. Habeeba *et al.* [88] applied MLP to detect Deepfake video with very little computing power by exploiting visual artifacts in the face region. As far as the performance concern in machine learning based Deepfake methods, it is observed that these approaches can achieve up to 98% accuracy in detecting Deepfakes. However, the performance entirely relies on the type of dataset, the selected features, and the alignment between the train and test sets. The study can obtain a higher result when the experiment uses a similar dataset by splitting it into a certain level of ratio, for example, 80% for a train set and 20% for a test set. The unrelated dataset drops the performance close to 50%, which is an arbitrary assumption.

Zhang *et al.* [33] introduced a GAN simulator that replicates collective GAN-image artifacts and feeds them as input to a classifier to identify them as Deepfake. Zhou *et al.* [34] proposed a network for extracting the standard features from RGB data, while [35] proposed a similar but generic resolution. Besides, in [36]_[38], researchers proposed a new detection framework based on physiological measurement, for example, Heartbeat.

At first, the deep learning-based method was proposed in [40] for Deepfake video detection. Two inception modules, (i) Meso-4 and (ii) MesoInception-4, were used to build their proposed network. In this technique, the mean squared error (MSE) between the actual and expected labels is used as the loss function for training. An enhancement of Meso-4 has been proposed in [41].

Disadvantages

High dimensional features are not preserved as a binary coded structure.

The information is not stored in a permission-based Blockchain, which gives the owner control over its contents.

III. PROPOSED SYSTEM

_ We perform a comprehensive survey on existing literature in the Deepfake domain. We report current tools, techniques, and datasets for Deepfake detection-related research by posing some research questions.

_ We introduce a taxonomy that classifies Deepfake detection techniques in four categories with an overview of different categories and related features, which is novel and the first of its kind.

_ We conduct an in-depth analysis of the primary studies' experimental evidence. Also, we evaluate the performance of various Deepfake detection methods using different measurement metrics.

_ We highlight a few observations and deliver some guidelines on Deepfake detection that might help future research and practices in this spectrum.

Advantages

Presents a generic framework based on Blockchain technology by setting up a proof of digital content's authenticity to its trusted source. Presents the proposed solution's architecture and design details to control and administrate the interactions and transactions among participants. Integrates the critical features of IPFS [114]-based decentralized storage ability to Blockchain-based Ethereum Name service.

IV. MODULES

Service Provider

In this module, the Service Provider has to login by using valid user name and password. After login successful he can do some operations such as Login, Train & Test Data Sets, View Videos Datasets Trained and Tested Accuracy in Bar Chart, View Videos Datasets Trained and Tested Accuracy Results, View Deep fake Prediction Type, View Deep fake Prediction Ratio, Download Predicted Data Sets, View Deep fake Prediction Ratio Results, View All Remote Users.

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the user’s details such as, user name, email, address and admin authorizes the users.

Remote User

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like register and login, predict deepfake detection type,view your profile.

V.CONCLUSION

This SLR presents various state-of-the-art methods for detecting Deep fake published in 112 studies from the beginning of 2018 to the end of 2020. We present basic techniques and discuss different detection models' efficacy in this work. We summarize the overall study as follows:

_ The deep learning-based methods are widely used in detecting Deep fake.

_ In the experiments, the FF++ dataset occupies the largest proportion.

_ The deep learning (mainly CNN) models hold a significant percentage of all the models.

_ The most widely used performance metric is detection accuracy.

_ The experimental results demonstrate that deep learning techniques are effective in detecting Deep fake. Further, it can be stated that, in general, the deep learning models outperform the non-deep learning models. With the rapid progress in underlying multimedia technology and the proliferation of tools and applications, Deep fake detection still faces many challenges. We hope this SLR provides a valuable resource for the research community in developing effective detection methods and countermeasures.

VI.REFERENCES

[1] *FaceApp*. Accessed: Jan. 4, 2021. [Online]. Available: https://www. faceapp.com/

[2] *FakeApp*. Accessed: Jan. 4, 2021. [Online]. Available: https://www. fakeapp.org/

[3] G. Oberoi. *Exploring DeepFakes*. Accessed: Jan. 4, 2021. [Online]. Available: https://goberoi.com/exploring-deepfakes-20c9947c22d9

[4] J. Hui. *How Deep Learning Fakes Videos (Deepfake) and How to Detect it*. Accessed: Jan. 4, 2021. [Online]. Available:

https://medium. com/how-deep-learning-fakes-videos-deepfakes-and-how-to-detect-itc0b50fbf7cb9

[5] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, ``Generative adversarial nets,'' in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 2014, pp. 2672_2680.

[6] G. Patrini, F. Cavalli, and H. Ajder, ``The state of deepfakes: Reality under attack,'' Deeptrace B.V., Amsterdam, The Netherlands, Annu. Rep. v.2.3., 2018. [Online]. Available: https://s3.eu-west-2.amazonaws.com/rep2018/2018-the-state-of-deepfakes.pdf[7] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, ``Face2Face: Real-time face capture and reenactment of RGB videos,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2387_2395, doi: 10.1109/CVPR.2016.262.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, ``Unpaired image-to-image translation using cycle-consistent adversarial networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Oct. 2017, pp. 2242_2251, doi: 10.1109/ICCV.2017.244.

[9] S. Suwajanakorn, S. M. Seitz, and I. K. Shlizerman, ``Synthesizing Obama: Learning lip sync from audio,'' *ACM Trans. Graph.*, vol. 36, no. 4, p. 95, 2017.

[10] L. Matsakis. *Arti_cial Intelligence is Now Fighting Fake Porn.* Accessed: Jan. 4, 2021. [Online]. Available: https://www.wired.com/story/gfycatarti _cial-intelligence-deepfakes/

[11] A. Rössler, D. Cozzolino, L.Verdoliva, C. Riess, J. Thies, and M. Nieÿner, ``FaceForensics: A large-scale video dataset for forgery detection in human faces,'' 2018, *arXiv:1803.09179*.

[12] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Niessner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt, ``Deep video portraits,'' *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1_14, Aug. 2018, doi: 10.1145/3197517.3201283.

[13] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, ``Everybody dance now,'' 2018, *arXiv:1808.07371*.

[14] T. Karras, S. Laine, and T. Aila, ``A style-based generator architecture for generative adversarial networks,'' in *Proc. IEEE/CVF Conf. Com- put. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4396_4405, doi: 10.1109/CVPR.2019.00453.

[15] D. Budgen and P. Brereton, ``Performing systematic literature reviews in software engineering,'' in *Proc. 28th Int. Conf. Softw. Eng.*, New York, NY, USA, May 2006, pp. 1051_1052, doi: 10.1145/1134285.1134500.

[16] Z. Stapic, E. G. Lopez, A. G. Cabot, L. M. Ortega, and V. Strahonja, ``Performing systematic literature review in software engineering,'' in *Proc. 23rd Central Eur. Conf. Inf. Intell. Syst. (CECIIS)*, Varazdin, Croatia, Sep. 2012, pp. 441_447.

[17] B. Kitchenham, ``Procedures for performing systematic reviews," Softw. Eng. Group; Nat. ICT Aust., Keele; Eversleigh, Keele Univ., Keele, U.K., Tech. Rep. TR/SE-0401; NICTA Tech. Rep. 0400011T.1, 2004.

[18] B. Kitchenham and S. Charters, ``Guidelines for performing systemati literature reviews in software engineering," Softw. Eng. Group; Keele Univ., Durham University Joint, Durham, U.K., Tech. Rep. EBSE-2007- 01, 2007.

[19] M. A. Babar and H. Zhang, ``Systematic literature reviews in software engineering: Preliminary results from interviews with researchers," in *Proc. 3rd Int. Symp. Empirical Softw. Eng. Meas.*, Lake Buena Vista, FL, USA, Oct. 2009, pp. 346_355, doi: 10.1109/ESEM.2009.5314235.

[20] H. Do, S. Elbaum, and G. Rothermel, ``Supporting controlled experimentation with testing techniques: An infrastructure and its potential impact," *Empirical Softw. Eng.*, vol. 10, no. 4, pp. 405_435, 2005.

[21] F. Matern, C. Riess, and M. Stamminger, ``Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 83_92, doi: 10.1109/WACVW.2019.00020.

[22] U. A. Ciftci, I. Demir, and L. Yin, ``FakeCatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 15, 2020, doi: 10.1109/TPAMI.2020.3009287.

[23] X. Li, Y. Lang, Y. Chen, X. Mao, Y. He, S. Wang, H. Xue, and Q. Lu, ``Sharp multiple instance learning for deepfake video detection," 2020, *arXiv:2008.04585*.

[24] L. Guarnera, O. Giudice, and S. Battiato, ``Fighting deepfake by exposing the convolutional traces on images," 2020, *arXiv:2008.04095*.

[25] M. Bonomi, C. Pasquini, and G. Boato, ``Dynamic texture analysis for detecting fake faces in video sequences," 2020, *arXiv:2007.15271*.

She received B.Tech Degree from Vizag institute of technology in 2013 and MASTER OF ENGINEERING (COMPUTER SCIENCE AND ENGINEERING) Degree from MVGR COLLEGE OF ENGINEERING in 2016. She is Member of ISEA. She is having Academic experience of more than 7 years. Her current area of research includes Machine Learning and Data Science. She published 6 papers in reputed International Journals. Attended Various Workshops and Faculty Development Programs.