Awareness to Deepfake: A resistance mechanism to Deepfake

Mohammad Faisal Bin Ahmed
Department of Informatics
Technical University of Munich
Munich, Germany
Email: faisalbin@outlook.com

M. Saef Ullah Miah
Faculty of Computing
University Malaysia Pahang
Pekan, Malaysia
Email: md.saefullah@gmail.com

Abhijit Bhowmik
Faculty of Science and Technology
American International University-Bangladesh
Dhaka, Bangladesh
Email: abhijit@aiub.edu

Juniada Binti Sulaiman
Faculty of Computing
University Malaysia Pahang
Pekan, Malaysia
Email: junaida@ump.edu.my

2021 International Congress of Advanced Technology and Engineering (ICOTEN) | 978-1-6654-1224-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICOTEN52080.2021.9493549

***Abstract*—The goal of this study is to find whether exposure to Deepfake videos makes people better at detecting Deepfake videos and whether it is a better strategy against fighting Deepfake. For this study a group of people from Bangladesh has volunteered. This group were exposed to a number of Deepfake videos and asked subsequent questions to verify improvement on their level of awareness and detection in context of Deepfake videos. This study has been performed in two phases, where second phase was performed to validate any generalization. The fake videos are tailored for the specific audience and where suited, are created from scratch. Finally, the results are analyzed, and the study's goals are inferred from the obtained data.**

***Index Terms*—Deepfake, Deepfake awareness, Artificial Intelligence, Cyber Crime, Deepfake study**

I. Introduction

Deepfake is a relatively new phenomenon in cyber-crime space. Deepfakes are videos or images generated by a form of artificial intelligence named deep learning, where the original subjects, specially faces of peoples are altered with other faces [1]. Deepfakes are a severe medium on spreading and creating fake news, hoax, defamation and pornography [2]. While developed countries can perceive it to an extent and are taking measures against it, this is hardly true for developing countries. As the study candidate country, Bangladesh is experiencing rapid digitalization due to its growing economy and increasing youth population. However, this makes the country an ideal target for cyber-criminal activities [3]. People there are generally not aware of recent technologies like Deepfake and its consequences [4]. While some might argue that detection is the best way to combat Deepfake, it's widely predicted that it will be impossible to detect Deepfake videos automatically in the future [5]. According to one of the creators of a prominent detection algorithm [6], detection is only a short term solution. So, despite being very accurate at the moment [7], this will not be enough. This study proposes another argument: awareness could be even more efficient in the fight against Deepfake. This argument can be established by studying the before and aftereffects of Deepfake exposure, in a developing country like Bangladesh.This study is designed to expose a group of volunteers to Deepfake videos who were previously unaware of the technology. After exposure, it would be analyzed what effects it leaves in the mind of those participants and how it changes their future online behaviors. Can they identify doctored videos? Can they realize its potential uses (and abuses?) Does Deepfake make them question the authenticity of real videos?

The sections of this paper organized as follows. Section II describes the methodology employed in this study. Results and findings of this study are discussed in section III, and with section IV the paper is concluded.

A. The Hypothesis

As the motive of this study, it is tested whether exposure to Deepfake videos make participants better at detecting Deepfake from a mix of real and fake videos. This study does not compare existing detection algorithms rather focuses on testing participants' detection ability in real life via awareness to Deepfake videos. Finally, some propositions are made in the context of Bangladeshi internet users and Deepfake as the use case of this study.

II. Methodology

This research is conducted in two phases to measure the effect of the exposure to Deepfake videos. Both parts consist of some videos and related questionnaires. It is conducted on two age groups: i) 18-24 years old and, ii) 24-30 years old. Educational background, technological prowess, privacy concerns, cyber -crime awareness, and prior history with other cyber-crimes are also accounted for all the participants. In the first phase of the study, participants are profiled and checked for their ability to detect Deepfake videos; then the test results are revealed to the participants individually and provided them more videos to watch. The videos had varying degrees of 'realness' to it. Some fake videos might be really easy to detect, others were very hard. After a few days, second

978-1-6654-1224-7/21/$31.00 ©2021 IEEE

phase of this study has been conducted which is checking their Deepfake detection ability by providing a mix of real and fake videos. The number of false positives and negatives are accounted on each part of the study. Finally, the study is concluded by finding out how the exposure had affected participants' perception of Deepfake: a) skepticism of other online contents, b) improvements to their own online habits, c) their Deepfake detection ability, and so on. These are dissected into the respective age groups, backgrounds, and gender. The study setup is loosely based on the 'Why Phishing Works' study hence some factors like gender, technology used are excluded from the questionnaires because there is generally no significant correlation of these to detection ability [8].

A. Compiling Deepfake Videos

Deepfake videos plays a vital role in this study and all the test questions are based on these videos. Three categories of videos are compiled for this study namely, 1. Local, 2. Politics 3. Entertainment. Most of the videos are collected from the internet apart from the local ones since it's almost nonexistent. There are a total of 20 videos for the whole study, however not all of them are fake videos. Some videos in the collection are un-doctored to introduce false positives and to check participator's detection ability. The videos are generally short in duration and the resolution are at least 720p. This is due to the fact that, most internet users in Bangladesh use smartphones and mobile data for consumption media contents [7]. Lower video quality might decrease their detection ability, and larger size videos take more time to render - making the experience difficult for the participants. To remove any discrepancy, all 20 videos are selected prior to the study and assigned to each phase by random. This is done to ensure the detection difficulty for each phase.

1) Making New Videos: Making Deepfake videos is not an easy task. Although there are many tools available now, selecting the right tool for this study, and then creating a data set for it is the hardest part. For this study, DeepFaceLab from "iperov" [9] is used. One of the reasons to choose this tool is its ease of integration with Google Colaboratory or "Colab" for short, (a product from Google Research) [10]; which is very useful if no CUDA [11] supported GPU are available locally for the computation tasks. Another reason is the online support - the tool has many detailed guides for diverse use cases.

Due to lack of CUDA equipped hardware, the videos are created using a Python Notebook running on Google Colab. Google Colab assigns GPU randomly and for this case, Nvidia T4 GPU was accommodated. For comparison, the same task was also run in a local machine equipped with AMD Radeon Pro 5300M GPU.

It is important to select two compatible videos for creating a Deepfake video. The videos should be detailed, should have a main subject and the subjects should appear for a long duration during the video. Usually, the more frames of faces there are, the better the model can be trained for it [12]. The general workflow for authoring a Deepfake video is as follows:

- Selecting source and target videos.
- Extracting face sets from these videos.
- Adjusting the face sets by sorting or removing.
- Training the model for these videos by selecting appropriate algorithm.
- Merging two videos according to trained model.
- Exporting final video with proper settings.

Initially this workflow was run simultaneously on both local machine and Google Colab. While the local machine completed 700 iterations, Google Colab was capped at 86000 iterations at the same time. Hence, the video from Google Colab was more accurate and exported for the task.

2) Selecting Existing Videos: Although there are many Deepfake videos available online, not all of them are suitable for this use case. For international videos, they are compiled from the available resources instead of making them by ourselves. As this approach is time saving, and there are already many good doctored videos available for example, around 3000 videos are provided by google and jigsaw which are available in public repository [13], [14]. However, when choosing the videos, following precautionary measures are taken:

- The videos shouldn't be widely known to be doctored.
- It's not obviously doctored.
- The videos should be interesting to our particular audience.

B. Participants Selection

As this study is taken place in Bangladesh as the case study country, all the participants are Bangladeshi citizens. For both studies, 100 participants were recruited - however, only 93 of them took both of the tests. They were recruited using social media and by person. The age groups range from 18-24 years and 24-30 years old. This was selected as this makes up the majority of Bangladesh's internet user groups [15], 87% of participants reside in Bangladesh; the rest resides in Germany and Malaysia.

C. The Study Setup

Like most two-phase study [16], the phases are done with the same participants. There are a 7 days buffer zone before the retest. This is done to clear out biases related to consecutive tests and to shorten the time span - making the study less verbose. The participants are provided with a Google Form link consists of the questions and the videos. After completion of each phase, there is a debriefing part that informs the participants about expected outcomes. To make it easier to detect how participants did in the tests, point system is introduced. This also has a side effect: making participants willing to do good during the tests. Point based questionnaires make participants more attentive and the tests less monotonous. The forms were sent to the participants by email or text messages. Each participants are allowed to response only once.

1) First Phase Setup: The first phase of the study is lengthier of the two. This phase consists of the following question types:

- Personal identifier

- Demographic questions
- Education and IT knowledge
- Internet usage pattern
- Prior cyber-crime exposure
- Answer confidence

By design, there are less questions in the first part of the study.

2) Second Phase Setup: The second phase of the study began exactly 7 days after the first phase. During this time, many participants expressed their interests to fake videos and some even forwarded their own findings about Deepfake. This indicates the first phase did indeed create awareness among many participants. This phase has more questions than its prior counterpart. The questions are more descriptive and contextual. For clarity and ease of understanding, Bengali translation are included in the questions. Same number of videos are selected for this phase. After completion, participants were shown how they did this time. However, how other participants did, or the status quo of the study are not disclosed to the participants at any point of the study.

Resources regarding this study like survey questions, Deepfake generator are available at the following git repository, https://github.com/ping543f/Deepfake.

III. Results and Discussions

The results from the surveys are dissected in two ways: first the answers to direct questions, and, the analysis from those answers. Since the surveys are conducted in Google Forms, the answers are easy to visualize. As for the latter, all the answers are exported to Google Sheets. The findings from both of the phases are merged, adjusted, and aptly trimmed. Finally, the key findings are surmised by comparing and tabulating.

A. First Phase Results

From the direct questions asked in the first phase, following findings are inferred:

- Around 45% participants use internet for more than 7 hours a day.
- 48% of the participants use internet only for social network and entertainment.
- 78% participants have never faced any sort of cyber-crime.
- 65% participants are totally unaware of Deepfake.

The last bit is important - this strengthen the fact that most participants are unaware of the phenomenon despite being a very active internet user. This is significant, because participants who already knew about Deepfake did much better in the first phase than participants who did not. The second phase of the study makes it more concrete. It's worth mentioning, young group generally did worse at detecting Deepfake correctly - a trend that continues in the second phase of the study.

B. Second Phase Results

The second phase of the study includes many contextual questions. From the direct answers, it is surmised:

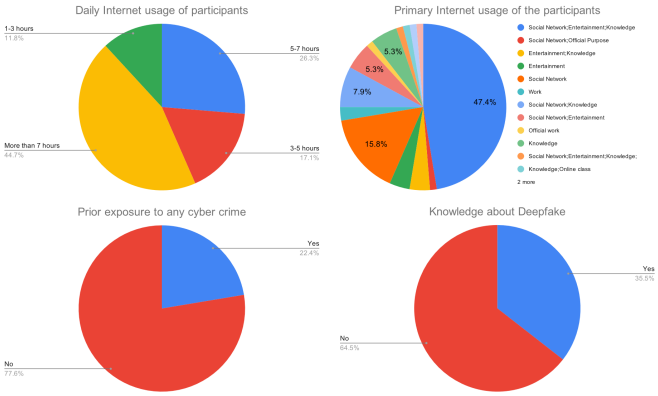

Fig. 1. Summary of first phase findings

- 60% of the participants said, exposure to Deepfake had made them more skeptical of video content on the internet and 22% are not sure and rest of the participants denied.
- Only 13% participants think that general population will able to distinguish between real and Deepfake videos and 57% of them think the opposite.
- Confidence level of the participants' have increased from those of phase 1.
- 70% participants said their detection ability improved after the initial phase where 20% of them are not sure about their improvement on detecting Deepfake videos.
- 82% said, they had improved their detection ability by learning about the Deepfake phenomenon after taking the phase 1 survey.

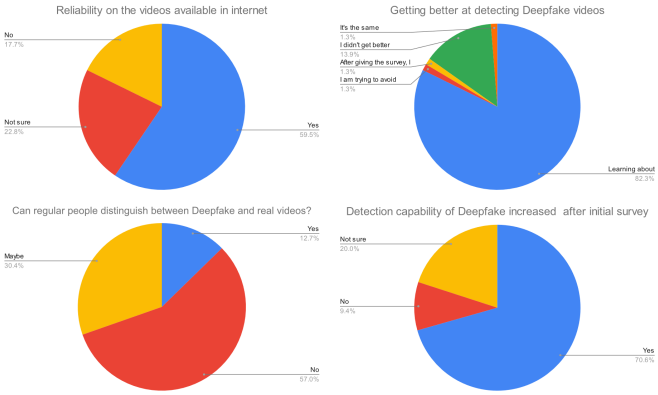

Fig. 2. Summary of second phase findings

The major takeaway from these findings is the last bit: a decisive number of participants said that exposure to Deepfake and consecutive learning experience have increased their ability to detect Deepfake videos correctly. This claim is sound, as majority of these participants did indeed perform better than the rest. Their total scores are well above the average which can be depicted from Figure 2.

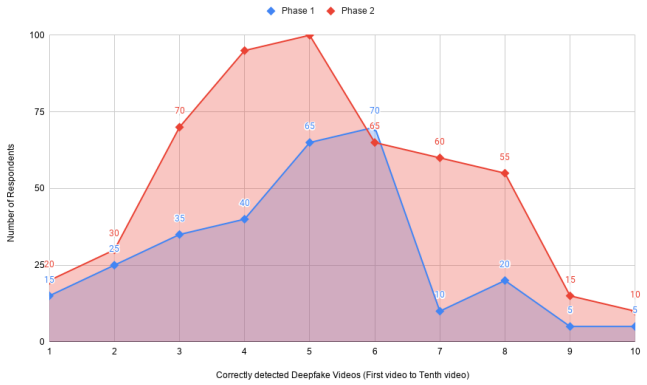

Fig. 3. Comparison of correctly detected Deepfake videos in phase 1 and phase 2

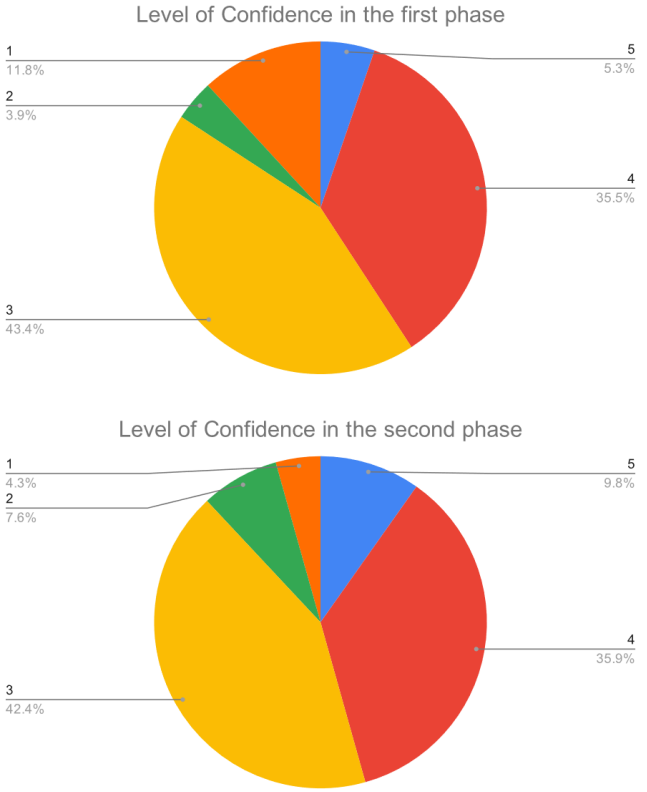

Fig. 4. Comparison of participants' confidence level in phase 1 and phase 2

C. Key Findings

By comparing and analyzing the outcomes from both results, we can draw the inference that,

- Around 71% participants improved their score in the second round while only 9% didn't and rest are not sure about their improvement. This can be depicted from Figure 2.
- Participants who learned about Deepfake after the initial phase, significantly improved their detection ability in the second phase. Figure 3 depicts the number of correctly detected Deepfake videos along with the number of participants.
- Confidence level of the participants have been increased in the second phase survey than phase 1 which is depicted in Figure 4.

These key findings bolster the argument that exposure to Deepfake does indeed increase people's detection ability. This can also be concluded that the initial exposure made them more suspicious of fake videos in general; as shown from the result, more skeptic, and experienced people do better than rest of the sample. Moreover, the detection rate increased across the board.

IV. Conclusion

The goal of this study was to find whether exposure to Deepfake videos makes people better at detecting them. Despite having a small sample size and other limitations, it can be derived from the results that exposure to Deepfake videos does increase people's ability to detect Deepfake videos. Even now, it might be a better strategy to fight Deepfake compared to automated detection.

A. Limitations

This study couldn't reach its full potential due to following drawbacks:

- Small sample size - a sample size of just 100 is very short to surmise anything conclusive. While everything was done to counter this, the results might still be not representative of the reality.
- Lack of hardware - while using Google Colab to make videos sort of works, it's limiting in terms of number of algorithm available and tweaks. with proper hardware, the quality and quantity of the videos will increase.
- Prior research - There is no relevant research literature published for Deepfake in the context of Bangladesh or any developing country as of our knowledge.

B. Final Words

While social networks are debating the ethical side of the issue - if unchecked, Deepfake may soon become a digital pandemic in developing countries. Although currently deployed detection tools are very accurate and work for the time being - they are destined to be obsolete at some point. When Deepfake videos will be virtually indistinguishable from the original, it is the people - who will decide the authenticity of the content. It is high time governments, especially of the countries like Bangladesh, take this seriously and start some awareness campaign to minimize its effect as much as possible.

Acknowledgment

We would like to express our gratitude to the participants who volunteered in the study, people who gave advise throughout the study, and people whose work this study is based on. Finally, We would like to thank the professor of the chair of

Cyber-trust (Department of Informatics, Technical University of Munich) Jens Grossklags for his support and vision.

REFERENCES

[1] I. Sample, "What are deepfakes – and how can you spot them?," The Guardian, 13-Jan-2020 [Online]. Available: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them. [Accessed: 24-Feb-2021]

[2] Westerlund, M. 2019. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review, 9(11): 40-53. http://doi.org/10.22215/timreview/1282

[3] S. Kundu, K. A. Islam, T. T. Jui, S. Rafi, M. A. Hossain and I. H. Chowdhury, "Cyber crime trend in Bangladesh, an analysis and ways out to combat the threat," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon-si Gangwon-do, Korea (South), 2018, pp. 474-480, doi: 10.23919/ICACT.2018.8323800

[4] I. Mahmud, T. Ramayah, M. M. H. Nayeem, S. M. Islam, and P. L. Gan. "Modelling cyber-crime protection behaviour among computer users in the context of Bangladesh". In: Design solutions for user-centric information systems. IGI Global, 2017, pp. 253–273.

[5] J. Vincent. Deepfake detection algorithms will never be enough. June 2019. url: https: / / www . theverge . com / 2019 / 6 / 27 / 18715235 / deepfake - detection - ai - algorithms - accuracy-will-they-ever-work.

[6] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li. "Protecting world leaders against deep fakes". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019, pp. 38–45.

[7] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. "FaceForen- sics++: Learning to Detect Manipulated Facial Images". In: International Conference on Computer Vision (ICCV). 2019

[8] R. Dhamija, J. D. Tygar, and M. Hearst. "Why phishing works". In: Proceedings of the SIGCHI conference on Human Factors in computing systems. 2006, pp. 581–590.

[9] Iperov. iperov/DeepFaceLab. Mar. 2020. url: https://github.com/iperov/DeepFaceLab.

[10] Google Colaboratory. url: https://colab.research.google.com/notebook.

[11] "About CUDA," NVIDIA Developer, 06-Mar-2014. [Online]. Available: https://developer.nvidia.com/about-cuda. [Accessed: 24-Feb-2021]

[12] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection: A Survey," arXiv:1909.11573 [cs, eess], Jul. 2020 [Online]. Available: http://arxiv.org/abs/1909.11573. [Accessed: 24-Feb-2021]

[13] "Contributing Data to Deepfake Detection Research," Google AI Blog. [Online]. Available: http://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. [Accessed: 24-Feb-2021]

[14] "ondyari/FaceForensics," GitHub. [Online]. Available: https://github.com/ondyari/FaceForensics. [Accessed: 24-Feb-2021]

[15] M. Mamun and M. D. Griffiths. "The assessment of internet addiction in Bangladesh: why are prevalence rates so different?" In: Asian journal of psychiatry (2019).

[16] R. Tao, D. Zeng, and D.-Y. Lin, "Optimal Designs of Two-Phase Studies," Journal of the American Statistical Association, vol. 115, no. 532, pp. 1946–1959, Oct. 2020, doi: 10.1080/01621459.2019.1671200.