# GODDS: The Global Online Deepfake Detection System

**Marco Postiglione[1], Julian Baldwin[1], Natalia Denisenko[1], Luke Fosdick[1], Chongyang Gao[1], Isabel Gortner[1], Chiara Pulice[1], Sarit Kraus[2], V. S. Subrahmanian[1]**

[1]Northwestern University, Evanston, Illinois, USA
[2]Bar-Ilan University, Ramat Gan, Israel
vss@northwestern.edu

## Abstract

Fake audios, videos, and images are now proliferating widely. We developed GODDS, the Global Online Deepfake Detection system, for a specific user community, namely journalists. GODDS leverages an ensemble of deepfake detectors, along with a human in the loop, to provide a deepfake report on each submitted video/image/audio or VIA artifact submitted to the system. To date, VIA artifacts submitted by over 50 journalists from outlets such as the New York Times, Wall Street Journal, CNN, Agence France Press, and others have been run through GODDS. Unlike other deepfake detection systems, GODDS doesn't just focus on the submitted artifact but automatically derives context about the subject of the VIA artifact. Because context is not always available on all subjects, GODDS focuses on alleged deepfakes of high profile individuals, organizations, and events, where there is likely to be considerable contextual information.

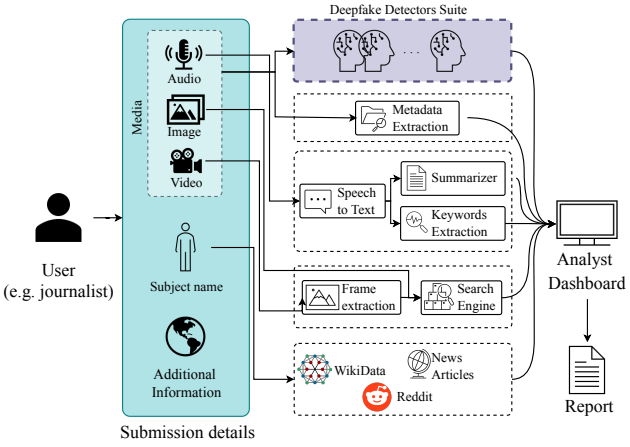



Figure 1: Architecture of the GODDS System

## Introduction

Deepfakes (Mirsky and Lee 2022; Verdoliva 2020) are now a growing threat to many stakeholders. A January 2024 deepfake audio of President Biden discouraging New Hampshire voters from voting in the primary election has led to steep fines on the perpetrators[1]. A video deepfake of Ukrainian President Zelenskyy telling his Army to lay down their weapons surfaced early in 2022[2]. Image deepfakes have depicted famous actresses in the nude[3]. There is now a critical need to rapidly detect and debunk such deepfakes, as well as alert platforms about deepfake content (Walker, Schiff, and Schiff 2024).

In response to this increasing threat, we have developed the Global Online Deepfake Detection System or GODDS. Though GODDS was made publicly available to journalists on July 8, 2024[4], earlier versions of the underlying algorithms and models have been used since September 2023.

GODDS differs from past work on deepfake detection in several respects:

1. First, GODDS does not just examine the VIA artifact submitted which is what most other deepfake detectors do (Ba et al. 2024; Zhang et al. 2024; Wang and Chow 2023; Tan et al. 2024). Instead, it automatically extracts the subject(s) portrayed in the artifact, then automatically derives context from open sources about the subject, and then applies multiple prediction algorithms. Because context is typically available only about public figures, significant organizations, and major events, this limits GODDS applicability to such subjects.
2. Because GODDS's primary user base is journalists who insist on high accuracy and detailed explanations, and because the main subjects of artifacts submitted to GODDS tend to be high profile (e.g. politicians, major world figures) where the cost of an error can be very high, GODDS includes a human in the loop review process in which analysts work in conjunction with the GODDS system to determine whether an VIA is a deepfake or not.

In the rest of this paper, we describe the architecture of GODDS, and our experiences with its use to satisfy the requests of over 50 journalists to date.

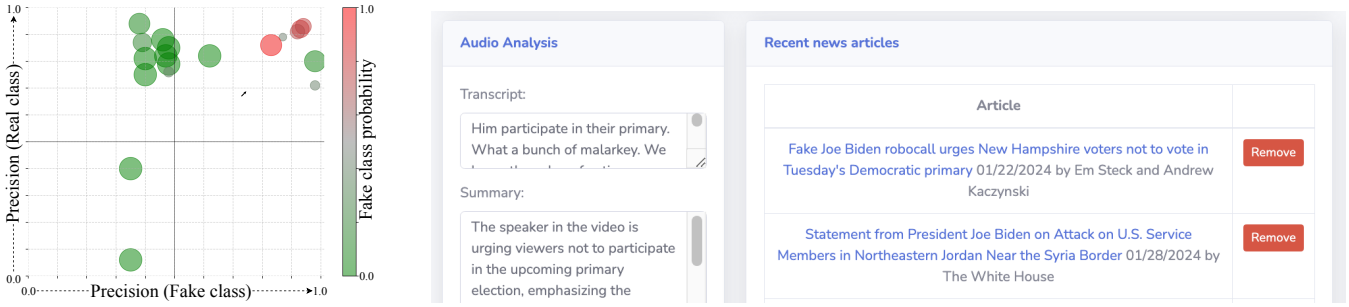


[1]https://www.npr.org/2024/05/23/nx-s1-4977582/fcc-ai-deepfake-robocall-biden-new-hampshire-political-operative

[2]https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia

[3]https://www.bbc.com/news/entertainment-arts-67726019

[4]https://godds.ads.northwestern.edu




(a) (b)

Figure 2: Screenshots of the GODDS System: (a) Performance and predictions chart for 20 audio deepfake detectors, (b) Partial context view.

## GODDS Architecture

Figure 1 shows the current architecture of GODDS. GODDS has three kinds of users: journalists, analysts, and administrators.

**Journalists** For legal reasons, only accredited journalists can set up accounts on GODDS. Once approved, a journalist may upload VIA artifacts. When they do so, they are also asked to provide any additional information they have (e.g. source of the video, any context they can provide). Some journalists provide additional information, others do not. Consider the infamous deepfake of President Biden calling voters in New Hampshire in January 2024, telling them not to come to the primary election. Once journalists hit the submit button on this audio, the data is stored in an underlying database and additional context about the VIA artifact is automatically gathered.

**Analysts** Figures 2a and 2b show a partial view of what our analysts would see when the Biden audio is uploaded. Figure 2a shows the predictions of 20 audio deepfake detectors. Each detector is represented as a dot. The x-axis shows the precision of the detector on the fake class and the y-axis shows the precision of the detector on the real class. Thus, the location of a dot tells us about the performance of the detector, while the color of the dot tells us what the detector is predicting on a continuous green to red scale. Dark green indicates high confidence the VIA artifact is likely real, dark red indicates high confidence that it is likely fake. We can see that several red dots that have high precision on the fake class of audios are shown toward the top right of the screen. But there are also some detectors predicting this artifact could be real.

Now imagine a CNN report based on the one detector that predicts this audio to be real. CNN would be severely criticized for this. This is why human trained analysts need to be in the loop. They can make an informed decision by looking at the artifact, examining the predictions made by different detectors on the authenticity of the artifact, looking at the context (part of the context is shown in Figure 2b, and assessing the performance of each detectors. Finally, they can mark up the artifact within the GODDS system, and produce a report which is automatically emailed to the journalist.

**Administrators** GODDS also has an administrative GUI using which our administrators can decide whether to approve/reject an account request, assign roles to users and add/remove deepfake detectors.

**Workflow** Once an artifact has been uploaded, GODDS initiates a thorough analysis utilizing a set of deepfake detectors. These methods include state-of-the-art algorithms (e.g. He et al. (2021); Cao et al. (2022); Kawa et al. (2023); Nguyen, Yamagishi, and Echizen (2019); Yan et al. (2023); Dang et al. (2020)) that are publicly available under licenses conducive to research (e.g. MIT, CC BY-NC 4.0), alongside novel techniques developed in our lab. We then extract metadata from the artifact, such as encoding methods and codec names, which can provide valuable indicators of potential deepfake characteristics. For media files containing audio, the system transcribes the spoken content into text with off the shelf tools. If the original language is not English, GODDS automatically generates an English translation. It also produces a summary and identifies keywords that describe the core themes of the audio with GPT-3.5 (OpenAI 2024). In cases where visual content is present, GODDS conducts a reverse image search using Google Images to locate visually related content. Additionally, GODDS enriches its analysis by collecting contextual information regarding the subject from diverse online sources, including news articles (WorldNewsAPI 2024; Webz 2024), Reddit posts (PRAW 2024), and WikiData (WikiData 2024). All this information is integrated into the analyst's dashboard, enabling analysts to conduct their assessments and submit reports to journalists.

The backend system for GODDS uses a LAMP stack, consisting of Linux, Apache, MySQL, and PHP, deployed on a Red Hat Enterprise Linux (RHEL) 8 environment. This architecture enables robust web service capabilities, facilitating efficient data handling and user interaction essential for the system's functionality.

## Users

GODDS has served over 50 journalists from a dozen countries: the United States, India, UK, Colombia, Turkey, Germany, Australia, Bangladesh, Bulgaria, France, Mexico, and Qatar. To date, all received feedback has been positive.

## References


Ba, Z.; Liu, Q.; Liu, Z.; Wu, S.; Lin, F.; Lu, L.; and Ren, K. 2024. Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 719–728. AAAI Press.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4113–4122.

Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; and Jain, A. K. 2020. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, 5781–5790.

He, Y.; Yu, N.; Keuper, M.; and Fritz, M. 2021. Beyond the Spectrum: Detecting Deepfakes via Re-Synthesis. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 2534–2541. ijcai.org.

Kawa, P.; Plata, M.; Czuba, M.; Szymanski, P.; and Syga, P. 2023. Improved DeepFake Detection Using Whisper Features. In Harte, N.; Carson-Berndsen, J.; and Jones, G., eds., *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, 4009–4013. ISCA.

Mirsky, Y.; and Lee, W. 2022. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.*, 54(1): 7:1–7:41.

Nguyen, H. H.; Yamagishi, J.; and Echizen, I. 2019. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, 2307–2311. IEEE.

OpenAI. 2024. GPT-3.5. https://openai.com. Accessed: 2024-12-14.

PRAW. 2024. PRAW: The Python Reddit API Wrapper. https://praw.readthedocs.io/en/stable/. Accessed: 2024-12-14.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. Frequency-Aware Deepfake Detection: Improving Generalizability through Frequency Space Domain Learning. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 5052–5060. AAAI Press.

Verdoliva, L. 2020. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process.*, 14(5): 910–932.

Walker, C. P.; Schiff, D. S.; and Schiff, K. J. 2024. Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 23053–23058. AAAI Press.

Wang, T.; and Chow, K. 2023. Noise Based Deepfake Detection via Multi-Head Relative-Interaction. In Williams, B.; Chen, Y.; and Neville, J., eds., *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 14548–14556. AAAI Press.

Webz. 2024. Webz News Retrieval API. https://webz.io. Accessed: 2024-12-14.

WikiData. 2024. WikiData. https://www.wikidata.org. Accessed: 2024-12-14.

WorldNewsAPI. 2024. WorldNewsAPI. https://worldnewsapi.com. Accessed: 2024-12-14.

Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhang, X.; Yi, J.; Wang, C.; Zhang, C. Y.; Zeng, S.; and Tao, J. 2024. What to Remember: Self-Adaptive Continual Learning for Audio Deepfake Detection. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 19569–19577. AAAI Press.