Creating More Meaningful Explanations for Audio Deepfake Detection

Bachelorarbeit

im Studiengang
Wirtschaftsinformatikim 6. Semester

vorgelegt von

Jaocb LaRock
Matr.-Nr.: 1667321
am 31.03.2025

Universität Siegen

Fakultät III: Wirtschaftswissenschaften, Wirtschaftsinformatik und Wirtschaftsrecht

Erstprüfer: Gunnar Stevens Gunnar.Stevens@uni-siegen.de

1 Abstract

Abstract filler text blablabla!

2 Introduction

As the use of generative methods for the creation of synthetic voices becomes more widespread, so grows the need for reliable and usable detection methods for the protection of the security of people and businesses alike. In particular, the rise of deepfake technology has led to concerns about its potential misuse in areas such as politics, entertainment, and national security by, for instance, potentially allowing malicious actors to create fake audio recordings that appear to be genuine statements made by public figures or to create genuine-seeming recordings of events that never happened, which could have significant consequences if used to spread misinformation, from defaming individuals to creating political tension [3, 1]. The field of Explainable AI (XAI) shows promise for producing useful and interpretable results from such models. Explainable AI refers to the ability of artificial intelligence systems, such as machine learning models and neural networks, to provide understandable and interpretable explanations for their decisions or predictions. This means that XAI systems can articulate why they made a particular decision, what factors influenced it, and how confident they are in their conclusion [2]. For the detection of audio deepfakes, this could mean that the model can justify its classification with feature-based evidence, allowing for verification of the result by the end user. Much of the existing research, further discussed in the next section, follows the path of producing a model with the best possible benchmarks against datasets of samples, without taking into account if the results can be usefully interpreted, while existing explorations into making audio deepfake detection explainable have produced results that often require a lot of background knowledge to meaningfully interpret With the context of the research discussed in the next section, I would like to use new combinations of features with a novel model architecture to generate

explanations for fake or real audio classification that are potentially more useful to their end user.

3 Related Work

4 Experiments

4.1 Dataset

4.2 Black-Box Model

4.2.1 Features

4.2.2 Model Architecture

4.3 Explanations

5 Results

5.1 Black-Box Model Results

5.2 Explanation Results

6 Discussion

7 Conclusion

References

[1] Marwan Albahar and Jameel Almalki. "DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW". In: . *Vol.* 22 (2005).

[2] Michael Hind. "Explaining explainable AI". In: *XRDS: Crossroads, The ACM Magazine for Students* 25.3 (Apr. 10, 2019), pp. 16–19. ISSN: 1528-4972, 1528-4980. DOI: 10.1145/3313096. URL: https://dl.acm.org/doi/10.1145/3313096 (visited on 06/05/2024).

[3] Namosha Veerasamy and Heloise Pieterse. "Rising Above Misinformation and Deepfakes". In: *International Conference on Cyber Warfare and Security* 17.1 (Mar. 2, 2022), pp. 340–348. ISSN: 2048-9889, 2048-9870. DOI: 10.34190/iccws.17.1.25. URL: https://papers.academic-conferences.org/index.php/iccws/article/view/25 (visited on 09/23/2024).