



SEMINARARBEIT

Titel der Arbeit

Deep Fake Erkennung: Entwicklung und Anwendung von
Methoden zur Erkennung von Deep Fakes in Video und
Audiomaterial

vorgelegt von
Michael Küchenmeister

Ingolstadt, 23.06.2024

Studiengang
Fakultät
Matrikelnummer
Prüfer:

Cloud Applications & Security Engineering
Informatik
00095870
Prof. Dr. Stefan Hahndel

Erklärung nach § 18 Abs. 4 Nr. 7 APO THI

Hiermit erkläre ich, Michael Küchenmeister, dass ich die vorliegende Seminararbeit selbstständig verfasst und noch nicht für anderweitige Prüfungszwecke vorgelegt habe. Ich habe keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt, sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet.

Ingolstadt, 23.06.2024
Ort, Datum

Küchenmeister, Michael
Unterschrift Michael Küchenmeister

Inhaltsverzeichnis

Erklärung nach § 18 Abs. 4 Nr. 7 APO THI	i
Tabellenverzeichnis	iii
Abbildungsverzeichnis	iv
1. Einleitung	1
2. Grundlagen zu Deepfakes	2
2.1. Deepfakes von Gesichtern	2
2.1.1. Face Replacement	3
2.1.2. Face Reenactment	3
2.1.3. Synthesierung und Bearbeitung von Gesichtsbildern	3
2.2. Deepfakes von Stimmen	3
2.2.1. Text-to-Speech	4
2.2.2. Voice Conversion	4
3. Methoden zur Erkennung von Deepfakes	5
3.1. Menschliche Wahrnehmung	5
3.2. Deep Learning basiert	6
3.3. Quellenverifizierung	7
4. Anwendung von Methoden zur Erkennung von Deepfakes	8
4.1. Menschliche Wahrnehmung	10
4.2. Deep Learning basiert	13
4.2.1. Funktionsweise der Deepfake Video Detection Modelle	13
4.2.2. Ergebnisse aus der Anwendung der Deepfake Video Detection Modelle	14
4.2.3. Funktionsweise des Deepfake Audio Detection Models	15
4.2.4. Ergebnisse aus der Anwendung des Deepfake Audio Detection Models	16
5. Fazit	17
Literaturverzeichnis	18
A. Anleitung zum Verwenden der KI-Modelle	20

Tabellenverzeichnis

4.1. Vorhersage für die Fälschung eines Videos	15
4.2. Vorhersage für die Echtheit einer Audio	16

Abbildungsverzeichnis

2.1. Veranschaulichung von Face Reenactment, Face Replacement, Face Editing und Face Synthesis [4]	2
4.1. Waveform der Audio LJ045-0087.wav	8
4.2. Waveform der Audio LJ045-0087_gen.wav	8
4.3. Waveform der Audio LJ050-0082.wav	8
4.4. Waveform der Audio LJ050-0082_gen.wav	9
4.5. Vorschaubild des Videos video_1.mp4	9
4.6. Vorschaubild des Videos video_1_gen.mp4	9
4.7. Vorschaubild des Videos video_2.mp4	10
4.8. Vorschaubild des Videos video_2_gen.mp4	10
4.9. Menschliche Analyse von video_1_gen.mp4	11
4.10. Menschliche Analyse von video_2_gen.mp4	12

1. Einleitung

Die rasante Entwicklung der künstlichen Intelligenz hat in den letzten Jahren zur Entstehung von sogenannten Deepfakes geführt. Das sind täuschend echt wirkende Fälschungen von Video- und Audiomaterial. Diese Technologie nutzt fortschrittliche Deep Learning-Algorithmen, um Gesichter in Videos zu manipulieren und Stimmen so zu verändern, dass sie authentisch erscheinen. Während Deepfakes faszinierende Anwendungen in der Filmindustrie und im Unterhaltungssektor finden, bergen sie gleichzeitig erhebliche Risiken für die Verbreitung von Desinformation, Betrug und Rufschädigung. Daher hat die Erkennung und Bekämpfung von Deepfakes sowohl in der Wissenschaft als auch in der Praxis an Bedeutung gewonnen [1, 2, 3].

Im Rahmen von fünf Kapiteln dieser Seminararbeit werden die Entwicklung und Anwendung von Methoden zur Erkennung von Deepfakes in Video- und Audiomaterial untersucht. Zunächst werden die Grundlagen zu Deepfakes erläutert, wobei zwischen der Manipulation von Gesichtern und Stimmen unterschieden wird. Im dritten Kapitel werden verschiedene Erkennungsmethoden vorgestellt, die von der menschlichen Wahrnehmung bis hin zu fortschrittlichen Deep-Learning-Modellen reichen. Anschließend werden die Anwendung dieser Methoden sowie ihre Effektivität in der Praxis analysiert. Abschließend fasst das fünfte Kapitel die Erkenntnisse zusammen und gibt einen Ausblick auf zukünftige Entwicklungen und Herausforderungen in diesem Bereich.

Die Untersuchung und Anwendung der in dieser Arbeit vorgestellten Methoden zur Deepfake-Erkennung sind nicht nur von akademischen Interessen, sondern auch von großer praktischer Relevanz. Sie tragen dazu bei, die Integrität digitaler Informationen zu wahren und bieten wertvolle Ansätze zur Bekämpfung der zunehmenden Bedrohung durch manipulierte Medieninhalte.

2. Grundlagen zu Deepfakes

Unter einem Deep Fake versteht man grundsätzlich einen von einer Künstlichen Intelligenz generierten Inhalt, der für den Menschen authentisch erscheint. Die Inhalte kommen dabei überwiegend in Form von Foto-, Video- oder Audiomaterial vor. Der Begriff „Deepfake“ setzt sich aus den Wörtern „Deep“ und „Fake“ zusammen. „Deep“ bezieht sich dabei auf einen mittels künstlicher Intelligenz generierten Inhalt und das Wort „Fake“ gibt an, dass die Authentizität dieses Inhaltes nicht gegeben ist, es sich also in diesem Fall um eine Fälschung handelt. Deepfakes lassen sich somit als realistisch wirkende, aber unechte Bild-, Video- und Audiomedien definieren, die durch Methoden der künstlichen Intelligenz insbesondere dem Deep Learning erstellt wurden. Mittlerweile gibt es eine Reihe an verschiedenen Ausprägungen von Deepfakes und deren Erzeugungsart [4, 5].

2.1. Deepfakes von Gesichtern

In den vergangenen Jahren wurden verschiedene KI-basierte Techniken zur Manipulation von Gesichtern in Videos entwickelt. Diese Methoden verfolgen unterschiedliche Ziele, darunter das Austauschen von Gesichtern in einem Video „Face Replacement“, die Kontrolle über die Mimik oder Kopfbewegungen einer Person in einem Video „Face Reenactment“, sowie die Synthesierung und Bearbeitung von Gesichtsbildern [6, 5].

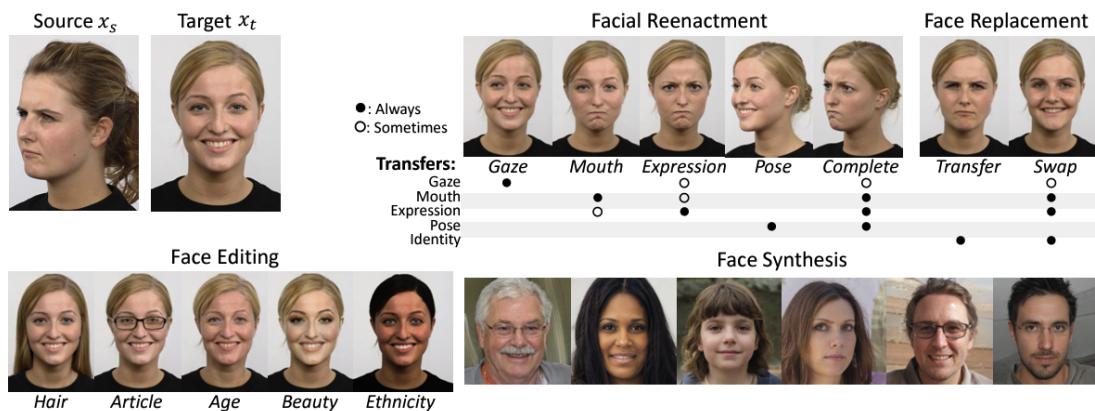


Abbildung 2.1.: Veranschaulichung von Face Reenactment, Face Replacement, Face Editing und Face Synthesis [4]

2. Grundlagen zu Deepfakes

2.1.1. Face Replacement

Bei diesen Verfahren wird das Gesicht einer Person x_t durch das Gesicht einer anderen Person x_s ersetzt. Beim Face Replacement wird außerdem zwischen Face-Transfer und Face-Swap unterschieden. Beim Face-Transfer wird der Inhalt aus x_t mit dem Inhalt aus x_s ersetzt und beim Face-Swap wird der von x_s zu x_t übertragende Inhalt durch x_t gesteuert. Face Transfer wird oft in der Modeindustrie verwendet, um beispielsweise einem Model virtuell unterschiedliche Kleidungsstücke anzuziehen oder Make-up aufzutragen. Face Swap ist die am weitesten verbreitete Form von Face Replacement und wird oft dazu genutzt, Memes oder Satire zu erzeugen, bei denen die Identität eines Schauspielers durch die einer bekannten Person ausgetauscht wird [7, 6].

2.1.2. Face Reenactment

Dieses Verfahren zielt darauf ab, die Mimik und Kopf-, Augen- oder Lippenbewegungen einer Person in einem Video nach Wunsch zu kontrollieren. Das Ziel hierbei ist es, das Gesicht einer Zielperson in einem gegebenen Video so zu manipulieren, dass es die Bewegungen und Ausdrücke einer anderen Quellperson widerspiegelt. Diese Technik ermöglicht es, Videos zu erzeugen, die visuell sehr überzeugend wirken und in denen eine Person Dinge sagt oder tut, die nicht der Realität entsprechen. Erreicht wird das Ganze durch die Erzeugung eines 3D-Modells des Gesichts einer Zielperson auf Basis eines vorgegebenen Videos [6, 7].

2.1.3. Synthesierung und Bearbeitung von Gesichtsbildern

Die Synthesierung umfasst die Erzeugung von fotorealistischen Bildern eines menschlichen Gesichtes, das in der Realität nicht existiert. Die Bearbeitung von Gesichtsbildern beinhaltet unter anderem die Retusche der Haut, die Veränderung der Haarfarbe oder aber auch das Verändern des Alters oder des Geschlechtes einer Person. In diesem Kontext kommen meist Generative Adversarial Networks (GANs) zum Einsatz. GANs basieren auf einem Fälscher (dem Generator) und einem Detektiv (dem Diskriminatator). Der Generator erstellt gefälschte Daten, während der Diskriminatator zwischen echten und gefälschten Daten zu unterscheiden lernt. Durch diesen „Wettstreit“ verbessern sich beide kontinuierlich, wodurch der Generator zunehmend realistischere Daten generieren kann [7, 6].

2.2. Deepfakes von Stimmen

Neben Deepfakes von Bild- und Videomaterial ist auch die Erstellung eines Deepfakes von Audiomaterial realisierbar. Auch hier gibt es verschiedene KI-basierte Methoden mit unterschiedlichen Zielen. Mit „Text-to-Speech“ Verfahren können geschriebene Texte in gesprochene Sprache umgewandelt werden und mit „Voice Conversion“ ist es möglich, die Stimme in einem Audio in die Stimme einer anderen Person umzuwandeln [6].

2. Grundlagen zu Deepfakes

2.2.1. Text-to-Speech

Bei Text-to-Speech-Verfahren spricht man von einem Prozess, welcher einen geschriebenen Text in gesprochene Sprache umwandelt und sich das daraus resultierende Audiosignal wie eine vorgegebene Zielperson anhört. Ein Text-to-Speech-System besteht aus 5 grundlegenden Komponenten. Bei der Textanalyse wird der Eingabetext analysiert und in einer Liste von einzelnen Wörtern organisiert. Die Textnormalisierung ist dafür zuständig, den Text in eine aussprechbare Form umzuwandeln. Die Phonetische Analyse umfasst den Prozess der Umwandlung von geschriebenem Text in eine phonetische Transkription. Dies bedeutet, dass der Text in eine Folge von Lautsymbolen übersetzt wird. Das Konzept der Prosodie beschreibt das Zusammenspiel von Betonungsmustern, Rhythmus und Intonation in der Sprache, wobei die prosodische Modellierung die Emotionen des Sprechers beschreibt. Intonation ist hingegen die Variation der Sprachmelodie beim Sprechen, um beispielsweise Freude auszudrücken oder Fragen zu stellen [8, 6].

2.2.2. Voice Conversion

Voice Conversion ist ein Prozess, bei dem die Sprecherstimme in eine andere Zielstimme umgewandelt wird, während die linguistischen Inhalte beibehalten werden. Durch den Einsatz von Deep Neural Networks (DNNs) und speziellen Architekturen wie Generative Adversarial Networks (GANs) oder Variational Autoencoders (VAEs) können komplexe nichtlineare Beziehungen zwischen den Sprachmerkmalen erfasst und realistischere Ergebnisse erzielt werden. Diese Deep-Learning-Modelle lernen direkt aus den Rohdaten und sind in der Lage, hochdimensionale Merkmalsräume zu modellieren, was zu einer verbesserten Qualität der konvertierten Stimmen führt [9, 6].

3. Methoden zur Erkennung von Deepfakes

In der heutigen digitalen Ära, in der multimediale Inhalte mit beispiellose Geschwindigkeit generiert und geteilt werden, ist die Fähigkeit, authentisches Material von sogenannten "Deepfakes" zu unterscheiden, von entscheidender Bedeutung. In diesem Kapitel werden drei grundlegende Methoden zur Erkennung von Deepfakes untersucht. Die menschliche Wahrnehmung, die sich auf die angeborene Fähigkeit des Menschen verlässt, Unstimmigkeiten zu erkennen. Deep Learning-basierte Ansätze, die auf künstlicher Intelligenz und maschinellem Lernen beruhen, um Muster zu identifizieren und zuletzt die Quellenverifizierung - ein Verfahren, das die Herkunft und Integrität der Daten überprüft.

3.1. Menschliche Wahrnehmung

Die menschliche Wahrnehmung spielt aktuell noch eine entscheidende Rolle bei der Identifizierung eines Deepfakes, da sie auf der intuitiven Fähigkeit basiert, Inkonsistenzen und Anomalien zu erkennen. Die Algorithmen, welche zur Erzeugung eines Deepfakes verwendet werden, sind momentan noch nicht perfekt und erzeugen teilweise deutliche Artefakte. Durch das Wissen um diese Artefakte lässt sich die Fälschungserkennung erheblich verbessern. Vor allem in Echtzeitanwendungen hat ein Angreifer nicht die Chance, das von Artefakten betroffene Material manuell zu bereinigen [6].

Typische Artefakte in Videomaterial [10, 11]:

- Unnatürliche Körperbewegungen oder -formen: Hierbei gilt zu beobachten, ob sich die abgebildete Person ruckartig bewegt, die Bewegungen von einem Bild zum nächsten passen und die Positionierung und die Proportionen von Kopf und Körper übereinstimmen.
- Merkwürdige Färbung: Hierbei ist zu beobachten, ob sich die Beleuchtung von einem Frame zum nächsten ändert, ob die Hautfarbe ungewöhnlich erscheint oder Schatten an Stellen zu erkennen sind, wo keine existieren sollten.
- Seltsame Augenbewegungen: Hierbei ist zu beobachten, ob sich die Augen auf unnatürliche Weise bewegen oder das Blinzeln eigenartig erscheint.
- Unrealistisch wirkende Mimik und Emotionen: Hierbei ist zu beobachten, ob die Mimik der Person realistisch erscheint und die dargestellten Emotionen zu den Gesichtsausdrücken der Person passen.

3. Methoden zur Erkennung von Deepfakes

- Unnatürlich wirkende Zähne/Haare: Hierbei ist zu beobachten, ob die Zähne künstlich erscheinen und die Haare sehr starr sind, also keine Bewegungen aufzeigen.
- Inkonsistente Geräusche oder Ton: Hierbei ist zu beobachten, ob die Lippenbewegungen nicht synchron mit der Audio sind, die Stimme robotisch oder monoton klingt, einzelne Wörter falsch ausgesprochen werden oder Verzögerungen beim Sprechen auftreten.
- Unscharfe Bereiche: Hierbei gilt zu beobachten, ob unnatürlich wirkende, zu scharfe, unscharfe oder verschobene Bereiche zu erkennen sind und die Helligkeit, Sättigung oder Schäfe des Hintergrundes mit dem Vordergrundes übereinstimmt.

Typische Artefakte in Audiomaterial [6]:

- Metallischer Sound: Hierbei ist zu prüfen, ob sich der Sound der Stimme metallisch oder robotisch bzw. verzögert anhört.
- Falsche Aussprache: Hierbei ist zu prüfen, ob in der Audio Wörter falsch ausgesprochen werden.
- Monotone Sprachausgabe: Hier gilt zu prüfen, ob das Audio-Signal sehr monoton hinsichtlich der Betonung einzelner Wörter ist.
- Falsche Sprechweise: Hierbei ist zu prüfen, ob spezifische sprachliche Charakteristika einer Zielperson fehlen, also sich Akzente oder Betonungen kontrastieren.
- Unnatürliche Geräusche: Hierbei ist darauf zu achten, ob unnatürliche Hintergrundgeräusche oder vollkommene Stille in der Audio auftritt.

3.2. Deep Learning basiert

Künstliche Intelligenz kann nicht nur zur Erzeugung eines Deepfakes, sondern auch zu deren Detektion genutzt werden. Diese dafür verwendeten KI-Detektoren basieren auf maschinellem Lernen und tiefen neuronalen Netzen, die darauf trainiert sind, gefälschte Videos sowie Audios zu identifizieren. Grundsätzlich kann man den Erkennungsprozess von Deepfakes mittels KI in die folgenden vier Schritte untergliedern [4, 12, 13, 14].

1. Training der KI-Modelle: Die Modelle werden mit großen Mengen an echten und gefälschten Audio- oder Videodateien trainiert. Während des Trainingsprozesses lernen sie charakteristische Merkmale von Deepfakes zu erkennen.
2. Merkmalsextraktion: Die in Schritt 1 gelernten Merkmale werden automatisch erkannt und extrahiert.
3. Klassifikation: Die extrahierten Merkmale werden anschließend mittels eines Klassifikators bewertet. Hierbei kommen sowohl für Audios als auch Videos oft CNNs (Convolutional Neural Networks) zum Einsatz.

3. Methoden zur Erkennung von Deepfakes

4. Feinabstimmung und Aktualisierung: Damit die Modelle zukünftig neue Deepfake Verfahren möglichst zuverlässig erkennen können, müssen diese kontinuierlich mit neuen Datensätzen aktualisiert werden.

3.3. Quellenverifizierung

Ein weiterer Weg, ein Deepfake zu erkennen ist die Quellenverifizierung bzw. eine gute Medienkompetenz. Meist können Deepfakes durch ein fundiertes Verständnis und kritisches Hinterfragen von Medieninhalten entlarvt werden. Im Folgenden werden einige Punkte aufgezählt, die dabei helfen können, einen Inhalt auf dessen Vertrauenswürdigkeit zu analysieren [15].

- Untersuchung von Nachrichtenquellen: Hierbei sollte die Herkunft der Nachricht geprüft (Medien, Blogs, Social-Media-Konten) und Aspekte wie Legalität, Verlässlichkeit und die Einhaltung von Standards für Genauigkeit, Ausgewogenheit und Fairness berücksichtigt werden. Außerdem sollte man auf seltsame oder unbekannte URLs oder Domainennamen achten und die gegebenenfalls bereitgestellten Informationen über den Eigentümer überprüfen.
- Analyse des Inhalts: Zeichen für gefälschte Inhalte können eine einseitige Darstellung von Fakten, Meinungen und Kommentaren sein, eine verzehrte Darstellung von realen Daten, der Gebrauch von sensationellen oder schockierenden Überschriften, die Verwendung von nicht überprüften Fotos/Videos und Verweise auf soziologische Daten ohne Angabe der Stichprobengröße, des Auftraggebers und der Geographie der Studien. Materialien, die von fachfremden Bloggern präsentiert werden, die sich zugleich nicht an traditionelle redaktionelle Standards halten, sollten ebenfalls sehr kritisch bewertet werden.
- Überprüfung der Informationen über den Autor: Eine verlässliche Autorschaft ist ein starkes Argument für die Authentizität und Zuverlässigkeit einer Nachricht, wobei zu beachten ist, dass Artikel, Forschungen und Berichte von Analysezentren, Unternehmen, Fonds oder Interessengruppen oft aus ideologischen Gründen finanziert werden.
- Analyse von Links: Das Fehlen von Links zu Informationsquellen in der Nachricht kann auf subjektive oder unbegründete Aussagen des Autors hinweisen, welche von der Realität abweichen können.
- Überprüfung der Aktualität einer Nachricht: Ein weiterer wichtiger Anhaltspunkt ist den Nachweis ihrer Relevanz, d.h. das Datum und die Uhrzeit der Veröffentlichung einer Nachricht, festzustellen. Außerdem kann geprüft werden, ob in dem gleichen Zeitraum ähnliche Informationen in anderen Quellen auftreten, wodurch veraltete Informationen identifiziert werden können.

4. Anwendung von Methoden zur Erkennung von Deepfakes

Dieses Kapitel beleuchtet die Anwendung der beiden in Kapitel 3 erläuterten Methoden „Menschliche Wahrnehmung“ und „Deep Learning basiert“. Eine Quellenanalyse ist mit ausschließlich Audio- und Videomaterial schwer möglich und benötigt eher den Kontext zur Veröffentlichung des Materials. Deshalb verzichtet diese Arbeit auf die Anwendung einer Quellenverifizierung. Für die Anwendung der beiden anderen Ansätze werden allerdings Audio- und Videomaterial benötigt. Auf Basis der im Kapitel 4.2 vtrainierten Modelle wird das Audiomaterial von [16] für die realen und von [17] für die synthetischen Datensätze verwendet. Die Datensätze für die Videodetektion stammen von der Deepfake Detection Challenge [18]. Aus diesen Datensammlungen wurden für die genauere Betrachtung jeweils 4 Audios und 4 Videos ausgewählt. Bei der Auswahl wurde berücksichtigt, dass immer ein originales sowie ein gefälschtes Video und Audio einer Person vorliegen, um einen genaueren Unterschied aufzuzeigen.



Abbildung 4.1.: Waveform der Audio LJ045-0087.wav



Abbildung 4.2.: Waveform der Audio LJ045-0087_gen.wav



Abbildung 4.3.: Waveform der Audio LJ050-0082.wav

4. Anwendung von Methoden zur Erkennung von Deepfakes



Abbildung 4.4.: Waveform der Audio LJ050-0082_gen.wav

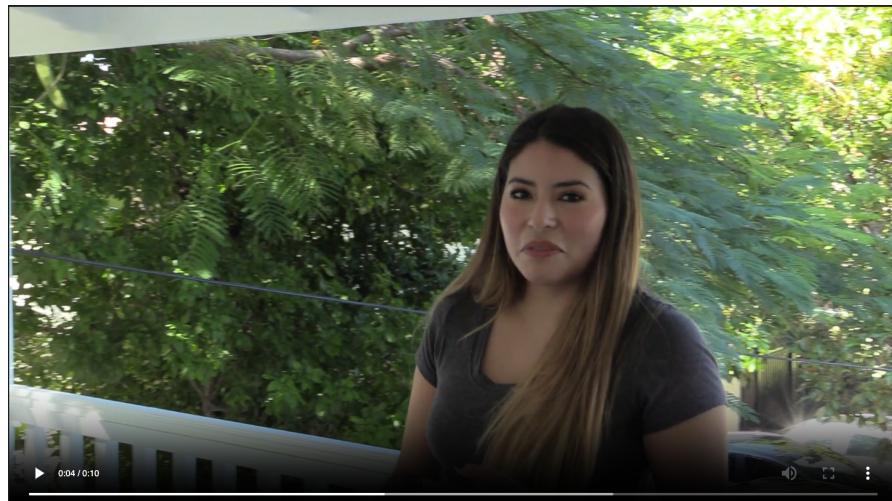


Abbildung 4.5.: Vorschaubild des Videos video_1.mp4



Abbildung 4.6.: Vorschaubild des Videos video_1_gen.mp4

4. Anwendung von Methoden zur Erkennung von Deepfakes



Abbildung 4.7.: Vorschaubild des Videos
video_2.mp4

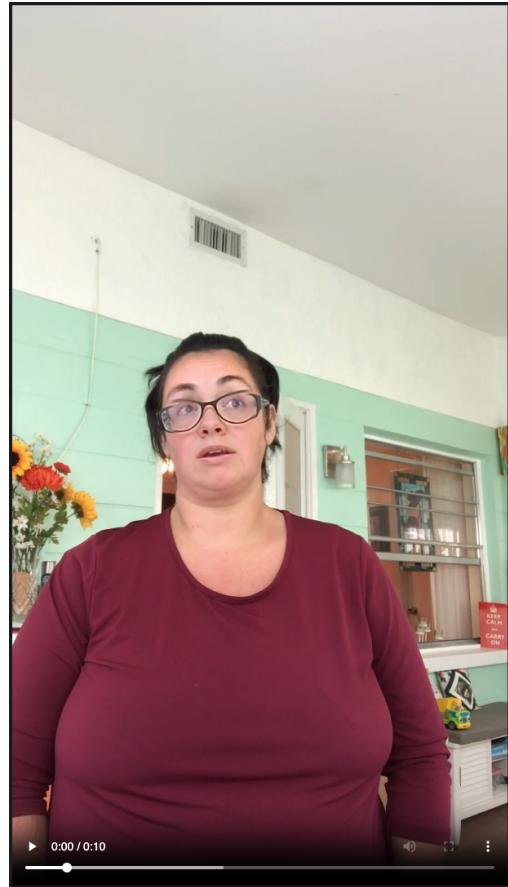


Abbildung 4.8.: Vorschaubild des Videos
video_2_gen.mp4

4.1. Menschliche Wahrnehmung

Um einem Deepfake mittels menschlicher Sinne zu erkennen, stützt sich diese Untersuchung auf die in Kapitel 3 erläuterten typischen Artefakte, die in Video- oder Audiomaterial vorkommen können. Die im Vorfeld ausgesuchten Audios und Videos (Abb. 4.1 - 4.8) wurden anschließend einzeln betrachtet und auf ebenjene untersucht.

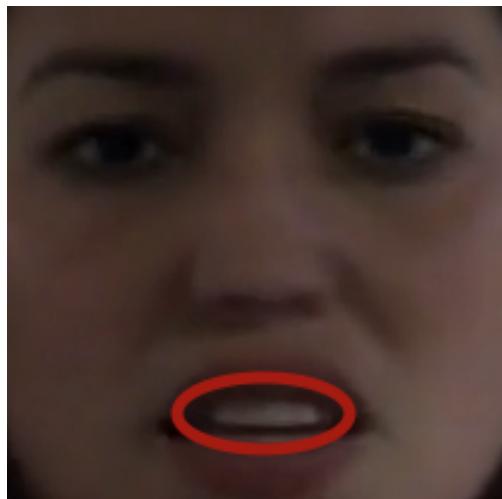
Untersuchung der Videos *video_1.mp4* und *video_1_gen.mp4*:

Beim Video *video_1.mp4* handelt es sich um ein reales Video der abgebildeten Person, während *video_1_gen.mp4* eine manipulierte Version ist. Nach eingehender mehrfacher Betrachtung beider Videos wurde zudem die Tonspur mit den Videoframes auf nicht zusammenpassende Auffälligkeiten untersucht sowie die Videoframes einzeln analysiert.

Beim Video *video_1.mp4* sind keine Auffälligkeiten aus Kapitel 3 erkennbar, weswegen es authentisch und realistisch erscheint.

4. Anwendung von Methoden zur Erkennung von Deepfakes

In *video_1_gen.mp4* sind hingegen die drei Artefakte „Merkwürdige Färbung“, „Unnatürlich wirkende Zähne“ und „Unscharfe Bereiche“ auffindbar. Ersteres erkennt man an einer leichten wiederkehrenden Schattenbildung auf der Stirn der Person, welches als Flackern interpretiert werden kann. Allerdings kann dieses Flackern auch legitimer Form sein und von der Kamera verursacht worden sein. Ein weiteres, eher schwaches Aufälligkeitsmerkmal sind die etwas künstlich wirkenden Zähne. Vor allem im Obergebiss der Person wirkt es in manchen Frames so, als ob die kleinen Zahnzwischenräume fehlen (siehe Abbildung 4.9.a). Obendrein scheint es aufgrund der deutlich in Abbildung 4.9.b zu erkennenden verschobenen bzw. unscharfen Bereiche im Nasenbereich so, als besitze die Person nur ein Nasenloch. Diese Auffälligkeit lässt sich im gesamten Videoverlauf feststellen.



(a) Unnatürlich wirkende Zähne



(b) Unscharfer Bereich

Abbildung 4.9.: Menschliche Analyse von *video_1_gen.mp4*

Untersuchung der Videos *video_2.mp4* und *video_2_gen.mp4*:

Das echte Video *video_2.mp4* und das Deepfake *video_2_gen.mp4* wurden auch hier mehrfach betrachtet und auf die einzelnen Artefakte aus Kapitel 3 untersucht. Des weiteren wurden erneut die Tonspur mit den Videoframes abgeglichen sowie die Frames einzeln betrachtet.

Video *video_2.mp4* zeigte keine Anzeichen auf ein Deepfake und scheint somit authentisch und nicht manipuliert zu sein.

Bei *video_2_gen.mp4* sind wiederum zwei Artefakte, die auf ein Deepfake hindeuten, auffällig geworden. Insbesondere bei der rechten Augenbraue dieser Person lässt sich ein auffälliges, wiederkehrendes Flackern bzw. eine leichte Verfärbung feststellen. Es wirkt so, als ob sich die Dichte der Haare von Frame zu Frame ändert (siehe Abb. 4.10.a -

4. Anwendung von Methoden zur Erkennung von Deepfakes

4.10.b). Des weiteren sieht man in Frame 00:05 einen unscharfen Bereich am Rahmen der Brille (siehe Abb. 4.10.c), woran man sehr gut erkennen kann, dass eine Brille über das eigentliche Gestell der Person gelegt wurde, da der ursprüngliche Rahmen deutlich zum Vorschein kommt.

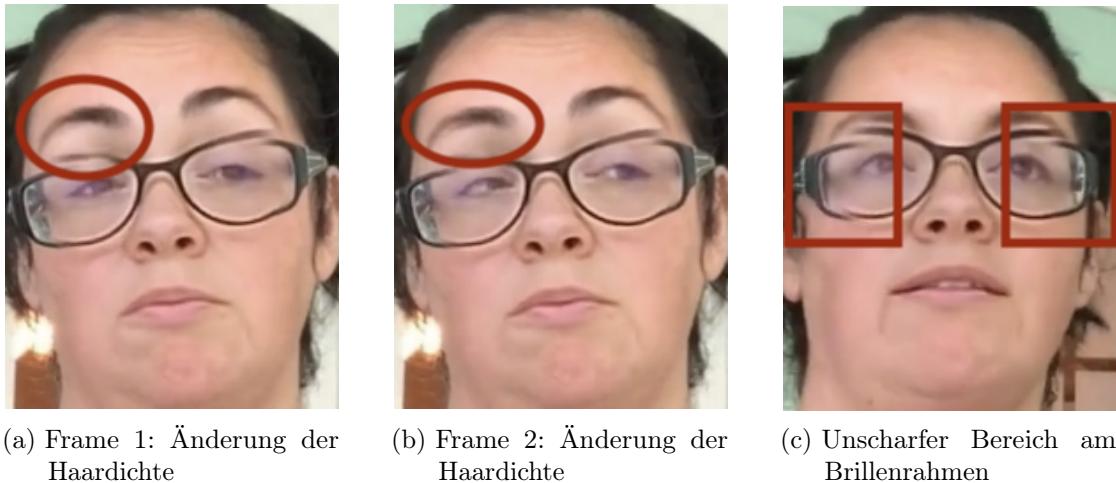


Abbildung 4.10.: Menschliche Analyse von video_2_gen.mp4

Untersuchung der Audios *LJ045-0087.wav* und *LJ045-0087_gen.wav*:

Bei der Audio *LJ045-0087.wav* handelt es sich um die Originalversion der echten Sprecherin, während Audio *LJ045-0087_gen.wav* eine Fälschung ist. Betrachtet man die Waveform der beiden Audios in den Abbildungen 4.1 und 4.2, kann man nur sehr schwer einen Unterschied erkennen. Bei genauerem Hinsehen kann man allerdings kleinste Unterschiede an einigen Stellen feststellen. Für die Anlayse auf die einzelnen Artefakte wurden beide Audios mittels Kopfhörern und eingeschaltetem Noise Canceling mehrfach angehört und untersucht.

Bei der Audio *LJ045-0087.wav* wurden keine besonderen Auffälligkeiten festgestellt, bis auf ein leichtes metallisches Aufschwingen bei der Aussprache des Konsonanten „S“. Dies könnte jedoch auch durch ein Übersteuern bei der Aufnahme mit dem Mikrofon verursacht worden sein oder an der Aussprache dieser Person liegen. Ansonsten hört sich die Audio authentisch an.

Für den direkten Vergleich wurde nun die gefälschte Audio *LJ045-0087_gen.wav* der Sprecherin mit den gleichen Voraussetzungen untersucht. Auch hier ist ein deutlich intensiverer metallischer Sound bei der Aussprache der Vokale - insbesondere eine zum Teil leicht verzögernde (laggy) oder roboterähnliche Sprechweise bei „O“ - wahrnehmbar. Weitere zutreffende Artefakte wurden nicht festgestellt. Aufgrund der deutlich zu hörenden metallischen und verzögerten Klänge, die vor allem im direkten Vergleich zum Original erfasst wurden, scheint diese Audio künstlich generiert worden zu sein.

4. Anwendung von Methoden zur Erkennung von Deepfakes

Untersuchung der Audios *LJ050-0082.wav* und *LJ050-0082_gen.wav*:

Auch bei diesen beiden Audios handelt es sich bei *LJ050-0082.wav* um das Original und bei *LJ050-0082_gen.wav* um die Fälschung. Bei diesem Beispiel wurden ebenfalls die gleichen Kopfhörer mit eingeschaltetem Noise Cancelling zur Analyse genutzt, wobei erneut in der Waveform (siehe Abb. 4.3 - Abb. 4.4) nur kleinste Unterschiede der beiden Audios erkennbar wurden.

Ähnlich der Audio aus dem vorherigen Beispiel konnten beim Originalaudio *LJ050-0082.wav* gleichermaßen keine außergewöhnlichen Auffälligkeiten festgestellt werden. Bei der Aussprache des Konsonanten „S“ konnte jedoch wiederum ein leicht zu hörender metallisch klingender Sound registriert werden. Stellt man aber *LJ050-0082.wav* und *LJ045-0087.wav* gegenüber, wird deutlich, dass die Sprecherin den Konsonant „S“ konstant so ausspricht. Weitere Auffälligkeiten wurden nicht registriert.

Analog zum ersten Beispiel trifft auch in dieser Audio nur der metallische Sound als erkennbares Artefakt zu. Dieser wird beispielsweise bei der Aussprache der Wörter „goes“, „harm“ und „embarrassment“ in den Sekunden 1,4 bis 1,8, 6,1 bis 6,4 und 6,6 bis 7,2 sehr gut wahrnehmbar. Die Stimme zeigt hier wieder leichte Verzögerungen (ist laggy) bzw. eine künstliche, roboterähnliche Aussprache der Wörter. Diese Anzeichen lassen darauf schließen, dass diese Audio wohlmöglich künstlich generiert/manipuliert worden ist.

4.2. Deep Learning basiert

Für die Erkennung von Deepfakes mittels künstlicher Intelligenz wird zunächst für die Analyse ein vortrainiertes KI-Modell benötigt. Bei der Recherche war vor allem wichtig, dass die Modelle bereits auf die Erkennung von Deepfakes in sowohl Video- als auch Audiomaterial vortrainiert sind, um einen langwiedrigen und ressourcenaufwändigen Trainingsprozess zu vermeinden. Nach einer ausgiebigen Recherche und dem ausprobierem mehrerer Modelle wurde letztendlich das *ShallowCNN_lfcc_I* Audio Deep Fake Detection Modell von Mark He Huang et. al. ausgewählt [19]. Für ein Modell zur Vorhersage von Deepfakes in Videomaterial wurde sich hingegen auf eine Kombination aus mehreren EfficinetNet B7 Modellen von Selim Seferbekov festgelegt [20]. Die zum Training genutzten Datensätze beruhen beim Audio Detection Model auf [16, 17] und beim Video Detection Model auf der Deep Fake Detection Challange [18].

4.2.1. Funktionsweise der Deepfake Video Detection Modelle

Die verwendeten Modelle stammen aus einem Contest für Algorithmen zur Erkennung von Deepfakes in Videomaterial. Umgesetzt wurde das Ganze mit einer frame-by-frame Klassifizierung unter der Verwendung von EfficientNet Modellen, welche für ihr performantes Verhalten in der Bilderkennungsaufgaben bekannt sind. Die Modelle stützen

4. Anwendung von Methoden zur Erkennung von Deepfakes

sich bei der Deepfake Erkennung nur auf die Gesichter, die in den Videos registriert werden können, nicht jedoch auf den Körper oder den Hintergrund. Die Gesichtserkennung wurde mittels dem Multi-Task Cascaded Convolutional Networks Face Detector und einer Extrahierung der Gesichter aus einzelnen Videoframes, umgesetzt. Die erkannten Gesichtsausschnitte wurden außerdem von jeder Seite um 30% beschnitten, um sicherzustellen, dass sich das Modell nur auf die wichtigen Bereiche (die Gesichter) bezieht, anstatt unnötige Hintergrundinformationen oder das gesamte Frame zu untersuchen. Für das Training wurden EfficientNet-Modelle verwendet, die zusätzlich mittels ImageNet und Noisy Student vorgenutzt wurden. Bei ImageNet handelt es sich um eine Bilddatenbank, die Millionen von annotierten Bildern tausender Kategorien enthält. Noisy Student ist eine Selbsttrainingsmethode, bei der ein Modell zunächst auf beschriftete Daten trainiert wurde und dann durch das Hinzufügen von Rauschen (z.B. Datenaugmentation) auf Basis einer Kombination aus beschrifteten und pseudobeschrifteten unbeschrifteten Daten weiter verbessert wird. Die input size wurde außerdem auf 380 x 380 Pixel festgelegt sowie Datenaugmentierung wie Farb- oder Helligkeitsanpassungen vorgenommen. Für das Training selbst wurden immer 32 Frames aus dem Video verwendet [20].

4.2.2. Ergebnisse aus der Anwendung der Deepfake Video Detection Modelle

Für das Ausführen der bereits vom Entwickler auf Deepfake Erkennung vorgenutzten KI-Modelle wurde der bereitgestellte Quellcode verwendet und durch eine Entfernung des Quellcodes für das Training, sowie kleiner weiterer Änderungen - wie das Nutzen der Apple M2 GPU - leicht abgewandelt. Eine Anleitung zur Verwendung des Quellcodes ist in Anhang A zu finden.

Die Vorhersage erfolgt durch einen mehrstufigen Prozess, der die Analyse einzelner Frames und die Aggregation der Vorhersagen über das gesamte Video hinweg umfasst. Jedes Video wird zunächst in einzelne Frames zerlegt und einzeln verarbeitet. Anschließend extrahiert das Modell aus jedem Frame die relevanten Merkmale, um zu bestimmen, ob dieser gefälscht oder echt ist. Für jeden Frame wird dazu ein Vertrauenswert (Confidence Score) berechnet, der die Wahrscheinlichkeit angibt, dass ein Frame ein Deepfake ist (je höher der Score desto wahrscheinlicher). Abschließend wird noch eine heuristische Durchschnittsbildung durchgeführt, damit eine Gesamtvorhersage für das Video getroffen werden kann[20].

- Wenn mehr als 11 Frames und mehr als 40% der Frames mit hoher Sicherheit als Deepfake erkannt werden, wird der Durchschnitt dieser Frames als entgültiger Vorhersagewert genommen.
- Wenn mehr als 90% der Frames mit geringer Sicherheit als Deepfake erkannt werden, wird der Durchschnitt dieser Frames als endgültiger Vorhersagewert genommen.
- In anderen Fällen wird einfach der Durchschnitt aller Frame-Vorhersagen genommen.

4. Anwendung von Methoden zur Erkennung von Deepfakes

Die Analyse wurde wieder auf den gleichen Datensätzen wie in Kapitel 4.1 - *video_1.mp4*, *video_1_gen.mp4*, *video_2.mp4* und *video_2_gen.mp4* - durchgeführt. Die Analysedauer dieser vier Videos betrug auf einem Macbook Pro mit Nutzung der M2 Pro GPU etwas mehr als eine Minute. Nach erfolgreichem Durchlaufen des Skripts sind die Ergebnisse in der Datei *submission.csv* zu finden und diese zeigt folgende Vorhersageergebnisse:

Filename	Prediction
video_1.mp4	0.01886
video_1_gen.mp4	0.9863
video_2.mp4	0.009094
video_2_gen.mp4	0.991

Tabelle 4.1.: Vorhersage für die Fälschung eines Videos

Die Ergebnisse zeigen, dass die beiden Fake Videos mit sehr hoher Wahrscheinlichkeit auch als solche von den Modellen erkannt werden. Beide realen Videos zeigen hingegen eine sehr geringe Wahrscheinlichkeit eines Fakes. Bei allen vier Videos belegen die Modelle somit sehr gute Ergebnisse. Zu beachten ist jedoch, dass die Modelle mit Datensätzen von unter anderem den in den Videos zu erkennenden Personen bzw. sehr ähnlichen Videos, die hauptsächlich auf Face Swapping und Face Reenactment Verfahren basieren, trainiert wurden. Bei der Analyse eines der Modelle noch unbekannten Videos oder einer unbekannten Person zeigen diese hingegen keine guten Ergebnisse. Bei dem Video *WDR_Fake.mp4*, wobei es sich um ein vom WDR mit einen synthetischen Charakter erstelltes Deepfake handelt, wird mit einer Wahrscheinlichkeit von **0.1066** vorhergesagt. Dieser Wert ist relativ gering und zeigt, dass KI-Modelle nur gute Vorhersagen treffen, wenn sie mittels Daten über die Person, wo eine Aussage getroffen werden soll, oder zumindest mit ähnlichen Videos, die auf den gleichen Erzeugungsverfahren basieren, trainiert worden sind.

4.2.3. Funktionsweise des Deepfake Audio Detection Models

Für ein vortrainiertes Modell für die Audio Deepfake Erkennung wurde das *Shallow-CNN_lfcc_I* von Mark He Huang et. al. ausgewählt [19]. Bei diesem Modell handelt es sich um ein flaches (shallow) Convolutional Neuronales Netzwerk (CNN). Flache CNNs verfügen im Vergleich zu tieferen Netzwerken über weniger Schichten, wodurch sie weniger rechenintensiv und schneller zu trainieren sind. Die Merkmalsextraktion der einzelnen Audiodateien erfolgt mittels Linear Frequency Cepstral Coefficients (LFCC), womit spektrale Eigenschaften aus der Audio erfasst werden. Diese beschreiben die Frequenzinhalte eines Audiosignals und deren Verteilung über die Zeit. Das Modell wurde auf reale und synthetisch generierte Datensätze einer bestimmten Sprecherin trainiert. Die synthetischen Datensätze wurden dabei mittels der MelGAN Technologie erzeugt. Für eine Verbesserung der Robustheit dieses Modells wurden Datenerweiterungstechniken wie Rauschaddition, Tonhöhenverschiebung und Zeitdehnung angewendet.

4. Anwendung von Methoden zur Erkennung von Deepfakes

4.2.4. Ergebnisse aus der Anwendung des Deepfake Audio Detection Models

Auch für diese Ausführung des vortrainierten KI-Modells wurde der von den Entwicklern bereitgestellte Quellcode verwendet und die unnötigen Codeeinheiten für das Training und der anderen im Paper entwickelten Modelle entfernt, sowie eine Vereinheitlichung mit dem Code für die Videoerkennung durchgeführt. Eine Anleitung zur Verwendung des Quellcodes ist in Anhang A zu finden.

Für die Vorhersage erfolgt zunächst die LFCC-Berechnung (Merkmalsextraktion) der Audiodatei. Anschließend werden die extrahierten LFCC-Merkmale normalisiert, um einen konsistenten Input für das Modell sicherzustellen. Die normalisierten Merkmale werden dann in das trainierte ShallowCNN-Modell eingespeist und von diesem verarbeitet. Das Modell verarbeitet diese Merkmale durch seine Faltungsschichten, die verschiedene Muster und Merkmale erkennen, die auf ein echts oder gefälschtes Audio hinweisen. Das Modell liefert abschließend einen Wahrscheinlichkeitswert, welcher angibt, ob es sich bei dieser Audio um ein Original handelt [19].

Für die Analyse wurden wieder die gleichen Datensätze wie in Kapitel 4.1 verwendet (*LJ045-0087.wav*, *LJ045-0087_gen.wav*, *LJ050-0082.wav* und *LJ050-0082_gen.wav*). Die Berechnungsdauer der Analyse betrug mit diesem Modell auf einem Macbook Pro mit der M2 CPU circa 5 Sekunden. Die Wahrscheinlichkeiten zur Vorhersage sind nach erfolgreichem Durchlaufen des Skripts in der Datei *submission.csv* zu finden.

Filename	Prediction
<i>LJ045-0087.wav</i>	0.9996809959411621
<i>LJ045-0087_gen.wav</i>	1.0657862503649085e-06
<i>LJ050-0082.wav</i>	0.9999948740005493
<i>LJ050-0082_gen.wav</i>	1.7916143406182528e-05

Tabelle 4.2.: Vorhersage für die Echtheit einer Audio

Diese Ergebnisse zeigen, dass sowohl *LJ045-0087.wav* als auch *LJ050-0082.wav* mit einer sehr hohen Wahrscheinlichkeit als reale Audios analysiert wurden. Die beiden synthetisch erzeugten Audios *LJ045-0087_gen.wav* und *LJ050-0082_gen.wav* zeigen hingegen eine sehr geringe Wahrscheinlichkeit, die gegen Null verläuft, dass es sich dabei um reale Audios handelt. Somit kann festgehalten werden, dass dieses Deepfake Audio Detection Model sehr gute Ergebnisse liefert. Für diese Ergebnisse ist es jedoch wichtig, dass das Modell mit Datensätzen trainiert wurde, die Stimmen der Person enthalten, über die eine Vorhersage getroffen werden soll. Gibt man diesem vortrainierten Modell nämlich beispielsweise eine Audio, auf dessen Stimme das Modell nicht trainiert wurde (*Frauenhofer_gen.wav*), liefert dies eine Wahrscheinlichkeit von **1,0**, dass es sich dabei um eine reale Audio handelt, wobei die Audio ein fake ist.

5. Fazit

Diese Arbeit hat drei mögliche Methoden zur Erkennung von Deepfakes in Video- und Audiomaterial dargelegt. Für eine genauere Untersuchung und Analyse wurden jedoch nur die beiden Methoden *Menschliche Wahrnehmung* und *Deep-Learning-basiert* angewendet.

Mittels menschlicher Wahrnehmung wurden typische Artefakte in Audio- und Videodateien identifiziert, die auf Deepfakes hinweisen. Insbesondere wurde festgestellt, dass metallische Klänge und Verzögerungen in gefälschten Audios sowie unscharfe Bereiche, merkwürdige Färbungen und unnatürlich wirkende Zähne in manipulierten Videos Indikatoren für Deepfakes sein können. Durch den direkten Vergleich von echten und gefälschten Medien konnten diese Artefakte erfolgreich identifiziert und analysiert werden.

Die Anwendung von Deep-Learning-Modellen zur Erkennung von Deepfakes hat ebenfalls vielversprechende Ergebnisse gezeigt. Die verwendeten Modelle für die Audiodetektion als auch für die Videodetektion konnten mit hoher Genauigkeit identifizieren, ob es sich bei den vorliegenden Dateien um Fälschungen oder Originale handelt. Beide Ansätze erzielten hohe Vorhersagegenauigkeiten, wenn sie auf bekannte Datensätze und Personenmerkmale trainiert wurden. Es wurde allerdings auch deutlich, dass die Modelle Schwierigkeiten haben, Deepfakes mit für das Modell noch unbekannten Personenmerkmalen oder Deepfakes, die mit anderen Techniken erstellt wurden, zu erkennen.

Zusammenfassend zeigt diese Arbeit, dass sowohl menschliche Sinne als auch KI-Modelle wirksame Methoden zur Erkennung von Deepfakes bieten, wobei die besten Ergebnisse durch eine Kombination aller in Kapitel 3 erwähnten Methoden erzielt werden können. Die Grenzen der KI-Modelle hinsichtlich der Generalisierbarkeit auf neue und unbekannte Deepfakes unterstreichen die Notwendigkeit kontinuierlicher Weiterentwicklung und Anpassung der Erkennungstechnologien.

Literaturverzeichnis

- [1] Hany Farid. "Creating, Using, Misusing, and Detecting Deep Fakes". In: *Journal of Online Trust and Safety* 1.4 (Sep. 2022). DOI: 10.54501/jots.v1i4.56. URL: <https://tsjournal.org/index.php/jots/article/view/56>.
- [2] Abdulqader Almars. "Deepfakes Detection Techniques Using Deep Learning: A Survey". In: *Journal of Computer and Communications* 09 (Jan. 2021), S. 20–35. DOI: 10.4236/jcc.2021.95003.
- [3] Bianca Steffes und Anna Zichler. "Deepfakes in Videoverhandlungen vor Gericht". In: *Datenschutz und Datensicherheit - DuD* 48 (März 2024), S. 158–163. DOI: 10.1007/s11623-023-1899-1.
- [4] Yisroel Mirsky und Wenke Lee. "The Creation and Detection of Deepfakes: A Survey". In: *ACM Computing Surveys* 54.1 (Jan. 2021), S. 1–41. ISSN: 1557-7341. DOI: 10.1145/3425780. URL: <http://dx.doi.org/10.1145/3425780>.
- [5] Tao Zhang. "Deepfake generation and detection, a survey". In: *Multimedia Tools and Applications* 81 (Feb. 2022). DOI: 10.1007/s11042-021-11733-y.
- [6] Bundesamt für Sicherheit in der Informationstechnik (BSI). *Deepfakes - Gefahren und Gegenmaßnahmen*. URL: https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html. (aufgerufen am: 15.05.2024).
- [7] Yisroel Mirsky und Wenke Lee. "The Creation and Detection of Deepfakes: A Survey". In: *ACM Comput. Surv.* 54.1 (Jan. 2021). ISSN: 0360-0300. DOI: 10.1145/3425780. URL: <https://doi.org/10.1145/3425780>.
- [8] E.Chandra D.Sasirekha. "Text to Speech: A Simple Tutorial". In: *International Journal of Soft Computing and Engineering (IJSCE)* 2 (März 2012), S. 275–278. ISSN: 2231-2307.
- [9] Berrak Sisman u. a. *An Overview of Voice Conversion and its Challenges: From Statistical Modeling to Deep Learning*. 2020. arXiv: 2008.03648 [eess.AS].
- [10] 2024 Sosafe. *Wie Sie Deepfakes zielsicher erkennen*. URL: <https://sosafe-awareness.com/de/blog/wie-sie-deepfakes-zielsicher-erkennen/>. (aufgerufen am: 11.05.2024).
- [11] 2024 Lawpilots. *Deepfake: Das Phantom im Netz – Gefahren und Abwehrmöglichkeiten für Unternehmen*. URL: <https://lawpilots.com/de/blog/it-sicherheit/deepfake-phishing-neue-gefahren-fuer-unternehmen-und-organisationen/>. (aufgerufen am: 11.05.2024).

Literaturverzeichnis

- [12] Ruben Tolosana u. a. *DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection*. 2020. arXiv: 2001.00179 [cs.CV].
- [13] Gourav Gupta u. a. “A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods”. In: *Electronics* 13.1 (2024). ISSN: 2079-9292. DOI: 10.3390/electronics13010095. URL: <https://www.mdpi.com/2079-9292/13/1/95>.
- [14] Pavel Korshunov und Sébastien Marcel. “Vulnerability assessment and detection of Deepfake videos”. In: *2019 International Conference on Biometrics (ICB)*. 2019, S. 1–6. DOI: 10.1109/ICB45273.2019.8987375.
- [15] Vitalii Tyshchenko und Tetiana Muzhanova. “DISINFORMATION AND FAKE NEWS: FEATURES AND METHODS OF DETECTION ON THE INTERNET”. In: *Electronic Professional Scientific Journal «Cybersecurity: Education, Science, Technique»* 2.18 (Dez. 2022), S. 175–186. DOI: 10.28925/2663-4023.2022.18.175186. URL: <https://csecurity.kubg.edu.ua/index.php/journal/article/view/413>.
- [16] 2017 LibriVox. *The LJ Speech Dataset*. URL: <https://keithito.com/LJ-Speech-Dataset/>. (aufgerufen am: 11.05.2024).
- [17] Joel Frank und Lea Schönherr. *WaveFake: A data set to facilitate audio DeepFake detection*. Version 1.2.0. Zenodo, Nov. 2021. DOI: 10.5281/zenodo.5642694. URL: <https://doi.org/10.5281/zenodo.5642694>.
- [18] 2020 Kaggle. *Deepfake Detection Challenge*. URL: <https://www.kaggle.com/c/deepfake-detection-challenge/data>. (aufgerufen am: 11.05.2024).
- [19] Mark He Huang et al. *Audio Deep Fake Detection*. URL: <https://github.com/MarkHershey/AudioDeepFakeDetection?tab=readme-ov-file>. (aufgerufen am: 11.05.2024).
- [20] Selim Seferbekov. *DeepFake Detection (DFDC) Solution by @selimsef*. URL: https://github.com/selimsef/dfdc_deepfake_challenge. (aufgerufen am: 11.05.2024).

A. Anleitung zum Verwenden der KI-Modelle

Zum Verwenden der in der Arbeit untersuchten Deepfake Detection Modelle muss zunächst das Repository geklont werden:

```
git clone https://github.com/neverchange95/deep-fake-detection.git
```

Verwenden des Audio Detection Models:

1. Wechseln Sie in das Verzeichnis detect_audio: *cd detect_audio*.
2. Installieren Sie die Abhängigkeiten: *pip install -r requirements.txt*.
3. Führen Sie den Deepfake Audio Detector aus: *python predict_audios.py*.
4. Die resultierenden Ergebnisse werden in der CSV-Datei *submission.csv* gesichert.
5. Für weitere Informationen lesen Sie die *README.md*

Verwenden des Video Detection Models:

1. Wechseln Sie in das Verzeichnis detect_video: *cd detect_video*.
2. Laden Sie sich die benötigten vortrainierten Modelle herunter: *sh download_models.sh*
3. Installieren Sie die Abhängigkeiten: *pip install -r requirements.txt*
4. Führen Sie den Deepfake Video Detector aus: *python predict_videos.py*.
5. Die resultierenden Ergebnisse werden in der CSV-Datei *submission.csv* gesichert.
6. Für weitere Informationen lesen Sie die *README.md*