TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Cyber Security Engineering

Rashad Gafarli 201815IVSB

Analysis of AI-generated Content: Open-Source Techniques and Countermeasures

Bachelor's thesis

Supervisor: Fuad Budagov (MBA)

Tallinn 2023

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Küberturbe tehnoloogiad

Rashad Gafarli 201815IVSB

Tehisintellekti genereeritud sisu analüüs: avatud lähtekoodiga tehnikad ja vastumeetmed

Bakalaureusetöö

Juhendaja: Fuad Budagov (MBA)

Tallinn 2023

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Rashad Gafarli

15.05.2023

Abstract

In current day and age as of this paper being written, the prevalence and usage of AI technology and its derivatives are constantly increasing. While worries and concerns of using AI algorithms for propaganda, misinformation, and other harmful activities are present, many people are still woefully unaware of the ease of use and access the public have to software capable of using AI algorithms to change videos. These range from harmless applications such as video stabilization to more malicious ones that manipulate or falsify the content of videos.

This thesis will research the open-source software capable of generating and detecting content with AI modifications. The paper will also provide an overview of the various AI techniques that can be used to modify videos, such as generative adversarial networks (GANs) and deep learning-based algorithms in addition to the techniques use in detecting and mitigating AI-generated video content that can be done freely.

The outcome of this research will enable startups and individuals to able to incorporate defensive measures against malicious activity done with the use of AI which in turn would lead to the improved detection accuracy with better understanding of AI-generated content, enhanced trust and credibility, and protection from deception and fraud overall.

This thesis is written in English and is 43 pages long, including 6 chapters, 5 figures and 1 table.

List of abbreviations and terms

AI	Artificial Intelligence
A2V	Audio-to-Video
AI	Artificial Intelligence
AV	Audio-Visual
CEO	Chief Executive Officer
CSET	Center for Security and Emerging Technology
CPU	Central Processing Unit
Conv-LSTM	Convolutional Long Short-Term Memory
DNN	Deep Neural Network
FSGAN	FaceSwaping Generative Adversarial Network
GAN	Generative Adversarial Network
ICT	Information and Communication Technologies
IT	Information Technology
ML	Machine Learning
RNN	Recurrent Neural Network
VDR	Video Dialogue Replacement
RNN	Recurrent Neural Network
HMM	Hidden Markov Models
LSTM	Long- and Short-Term Memory

Table of contents

List of figures

List of tables

1 Introduction

1.1 Motivation

With the rise of artificial intelligence (AI), there has been a surge in the use of AI-generated content, including text, images, and videos. While AI-generated content has the potential to revolutionize various industries, it also poses significant risks, including the spread of misinformation, cyberattacks, and privacy violations. [1] A prominent example of this would be that during the early part of 2020, a bank manager in Hong Kong fell victim to a deepfake fraud, where he was duped into authorizing transfers amounting to $35 million. The fraud was carried out through a phone call made by an imposter who was able to mimic the voice of a director from a company with whom the bank manager had previously spoken. The imposter provided emails from the director and a lawyer confirming the transfer of funds, which convinced the bank manager to proceed with the transactions. This incident highlights the potential danger posed by deepfakes in financial transactions and emphasizes the importance of developing effective detection techniques and anti-deepfake measures to mitigate the risks associated with these sophisticated scams.

One could say this is a rare phenomenon but on the contrary, it is rather relevant and growing. “The number of deepfake videos online has been increasing at an estimated annual rate of about 900%” [2], VMware mentions that “Two out of three respondents in our report saw malicious deepfakes used as part of an attack, a 13% increase from last year, with email as the top delivery method.” [3]

This thesis aims to provide an in-depth analysis of the open-source generation and detection of content generated and modified by AI. Specifically, this thesis will review the open-source techniques and countermeasures for detecting AI-generated content and discuss their strengths and limitations. By analyzing the current techniques and countermeasures for generating, detecting, and mitigating the risks associated with AI-

content, this thesis will provide a comprehensive understanding of the ease of use and challenges posed by AI-content.

In conclusion, the main goal is to contribute to the ongoing discussion on the risks associated with AI-generated content by analyzing the current state-of-the-art techniques and countermeasures. This thesis offers a comprehensive understanding of the challenges posed by AI-generated media.

1.2 Problem Statement

AI-generated content has been used in various applications, including online advertising, entertainment, and social media. While such has the potential to revolutionize these industries, it also poses significant risks, such as the spread of misinformation, cyberattacks, and privacy violations. Therefore, there is a growing interest in developing techniques and countermeasures to detect and mitigate the risks associated with AI-generated content. These techniques can be based on metadata analysis, image analysis, or language analysis, among others. Additionally, there is a need to address the ethical, legal, and social implications of AI-generated content, including issues related to privacy, security, and bias.

Some examples of AI-generated content include chatbots, automated news articles, and deepfake videos. While they pose many potential benefits, including increased efficiency and lower costs, also comes with significant risks. For example, AI-generated content can be used to spread misinformation, manipulate public opinion, and perpetuate biases. Additionally, the use of AI-generated content can raise legal and ethical questions, such as who is responsible for content generated by an algorithm.

For simple tasks such as generating text, there are many open-source libraries and pre-trained models available that can be used by individuals with little programming experience. Generating more complex content, such as high-quality images or videos, may require more advanced knowledge and specialized tools, but it is very possible. Although the technology for generating AI-generated content is becoming more accessible and user-friendly, it still requires a certain level of expertise and resources. It is essential to note that while the ease of generating AI-generated content is increasing, so are the risks associated with its misuse. Studying the techniques and countermeasures

for AI-generated content is critical to understanding the challenges and risks associated with this technology and developing solutions to mitigate these risks.

1.3 Research goal and objectives

The aim of this research is:

- To explore the current open-source techniques used in the creation of AI-generated content, and to identify the potential usability and effectives of such generating high-quality AI-generated content in various domains, including text, images, and video.
- To explore the open-source countermeasures that can be utilized to detect the dissemination of malicious or harmful AI-generated content.

The objectives for this research are:

1. Review current openly available technology related to AI content.
2. Generate and display usage of AI-influenced media.
3. To conduct experimentative analysis to determine effectiveness.

1.4 Research Questions

This research will cover answers to the following questions:

1. What can an individual who decides to utilize AI tools do with immediate access to the internet and tools at hand?

2. What level of quality can one expect from Open-source AI content generation as well as the effectiveness of the detection algorithms able to be used?

2 Theoretical Background

AI-generated content refers to any type of content that has been created, modified, or manipulated by artificial intelligence algorithms. These algorithms can use various techniques, such as deep learning, natural language processing, and computer vision, to generate or modify content such as text, images, and videos.

However, creating AI-generated content that is convincing and realistic can require a significant amount of time, effort, and expertise. The process of training an AI model to generate content can involve collecting and processing large datasets, fine-tuning the model's parameters, and validating the results to ensure that they are accurate and realistic.

2.1 Overview of AI-Generated Content

In the following sections this paper will cover the definition and types of AI-generated content, as well as its applications and implications for cybersecurity. The use of AI-generated content raises many concerns in cybersecurity including the potential for AI-generated attacks, such as deepfakes, which can manipulate images and videos to create fake content that appears to be real. Additionally, the use of AI-generated content in phishing attacks can lead to more sophisticated and targeted attacks, increasing the risk of data breaches and other cyber threats. However, AI-generated content also has the potential to enhance threat detection and response through the use of AI-powered security systems and automated threat analysis. It is important for cybersecurity professionals to understand the risks and benefits of AI-generated content in order to develop effective defence strategies and ensure the safety and security of their organizations.

2.1.1 Applications and Implications of AI-Generated Content

Fake content in the form of images and videos using digital manipulation with artificial intelligence (AI) approaches has become widespread during the past few years. Deepfakes, a hybrid form of deep learning and fake material, include swapping the face of one human with that of a targeted person in a picture or video and making content to mislead people into believing the targeted person has said words that were said by another

person. Facial expression modification or face-swapping on images and videos is known as deepfake. [4] Deepfake videos where the face of one person is swapped with the face of another person have been regarded as a public concern and threat. Rapidly growing advanced technology has made it simple to make very realistic videos and images by replacing faces that make it extremely hard to find the manipulation traces. Face-swap apps such as 'FakeApp' and 'DeepFaceLab' make it easy for people to use them for malicious purposes by creating deepfakes for a variety of unethical purposes. With the first deepfake video appearing in 2017 when a Reddit user transposed celebrities' faces in porn videos, multiple techniques for generating and detecting deepfake videos have been developed.

The technology of deepfakes, although capable of constructive purposes such as filming and virtual reality applications, has the potential to be utilized for destructive purposes. [5,6,7] Manipulation of faces in photographs or films poses a significant threat to global security. Faces are fundamental to human interactions, biometric-based human authentication, and identity services. As a result, convincing changes in faces have the potential to undermine security applications and digital communications. Deepfake technology is used to create various types of videos such as humorous or pornographic videos featuring the voice and image of a person without any authorized use. [8] Deepfakes can serve several purposes, including creating fake pornographic videos of well-known people, spreading fake news, impersonating politicians, and committing financial fraud. [9] Two prominent examples were 'Maisy Kinsley' and 'Katie Jones'. Both were fake personas with profiles on LinkedIn and Twitter respectively, which were involved in espionage campaigns. [10] Although initially, deepfakes were aimed at politicians, actresses, leaders, entertainers, and comedians for making pornographic videos [16], their use for bullying, revenge porn, terrorist propaganda, blackmail, misleading information, and market manipulation poses a genuine threat. [3]

Thanks to the increasing use of social media platforms such as Instagram and Twitter, as well as the availability of high-tech mobile phone cameras, it has become easier to create and share videos and photos. The research "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" highlights the role of social media. [5] The authors note that any content, real or fake can go viral. There are more than 3.2 billion images and 720,000 hours of video shared online daily. [11]

As digital video recording and uploading have become increasingly easy, digital manipulation of media can have significant consequences depending on the information being changed. Using advanced techniques from deep learning technology, deepfakes can create hyper-realistic videos and fake pictures. Accompanying the widespread use of social media, deepfakes can instantly reach millions of people and pose a danger by spreading false news, hoaxes, and fraud. [12] Social media as one of the most significant avenues for information diffusion makes such content rather dangerous. Many sites have a policy for reducing the quality of the media they post allowing lower quality mistakes and defects to go hidden by the degradation of media quality resulting in bogus and misleading information to spread quickly and widely through social media channels, with the potential to affect millions of people. [13] The rise in popularity and reputation of videos necessitates the introduction of proper tools to authenticate media channels and news. In view of the easy access and availability of tools to disseminate false and misleading information using social media platforms, it is becoming increasingly difficult to verify the authenticity of the content. [14] Current issues are attributed to digital misleading information, also called disinformation, and represent information warfare where fake content is deliberately presented to alter people's opinions. [15,16]

Deepfake services are popular on underground forums and are frequently used to target online banking and digital finance verification. Criminals interested in these services often possess copies of victims' identification documents, but also require a video stream of the victim to create or steal accounts. These accounts can be used for malicious activities such as money laundering or other illicit financial transactions. Criminal attacks using verification tools and techniques have evolved alongside the development of modern technology and online chat systems, creating new methods for bypassing verification schemes. Celebrities, high-ranking government officials, and corporate figures are most at risk due to their high-resolution images and videos being readily available online. Social engineering scams using deepfake videos of public figures' faces and voices are already being proliferated. [17]

The following is a list of potential deepfake attacks:

- Impersonation scams in which fraudsters pose as an account manager and make video calls using deepfakes to request money transfers or top-ups in phone balances from the victim's friends and family.

- Identity theft and creation of fraudulent accounts in banks, financial institutions, and even government services, by using deepfakes to bypass identity verification services and submit copies of stolen identity documents. Criminals can use a victim's identity and bypass the video verification process to create accounts that can later be used in money laundering and other malicious activities.

- Account hijacking, in which deepfakes can be used to take over accounts that require identification using video calls, enabling criminals to withdraw or transfer funds. Financial institutions that use online video verification to enable certain features in their online banking applications are also at risk.

- Extortion, in which malicious actors can create deepfake videos to create more powerful extortion attacks, or even plant fake evidence to achieve their objectives.

- Disinformation campaigns, which can be used to manipulate public opinion or engage in schemes like pump-and-dump fraud. Deepfakes can be used to create messages from well-known persons that can have financial, political, and reputational repercussions.

- Hijacking of internet-of-things devices, such as those that use voice or face recognition, to gain unauthorized access to IT assets.

2.1.2 Definition and Types of AI-Generated Content

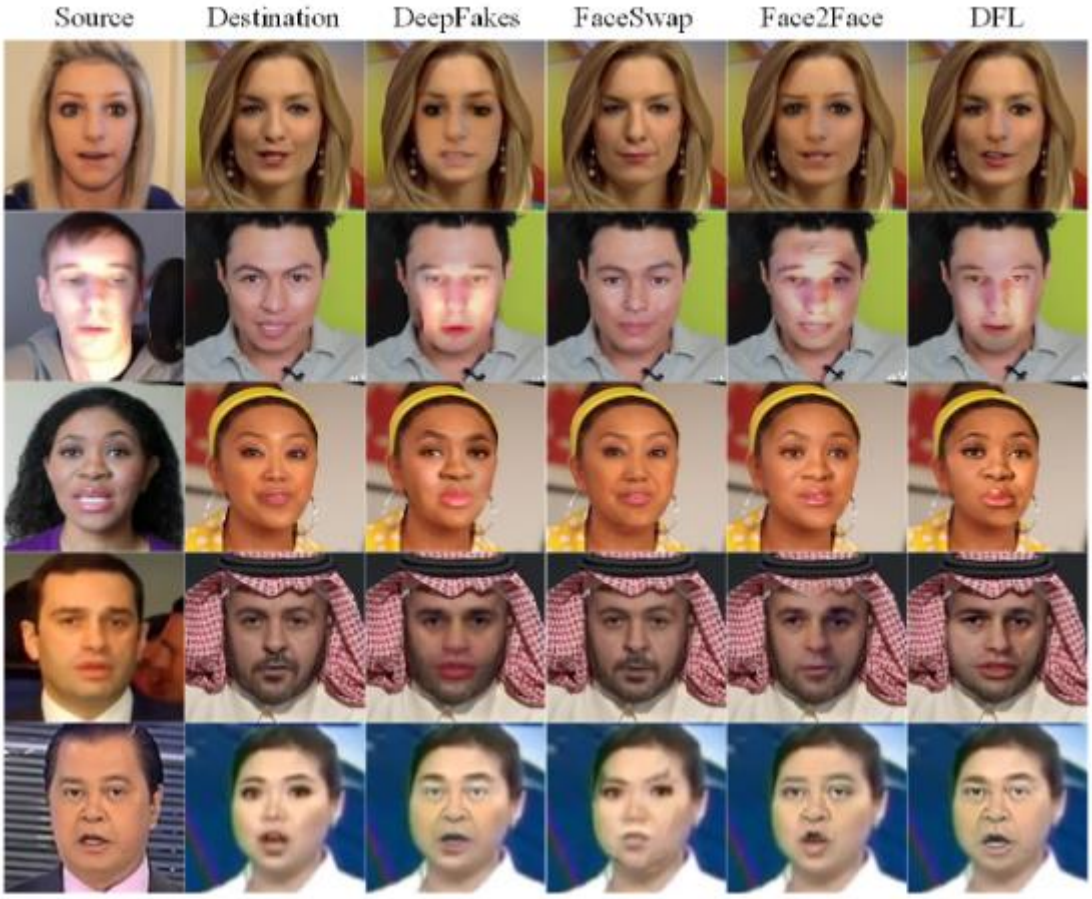

Figure 1 Faceforensics++ dataset example [18]

The authors of the paper "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking" note that the creation and editing of content in the past were typically time-consuming and required meticulous work. However, with the advent of Generative Deep Neural Network (DNN), the way content is synthesized and edited has changed significantly. One such software, DeepFake, which utilizes this approach, became publicly available in 2018 and has been widely used to create a large amount of fake content. [19] The authors assert that the proliferation of DeepFake technology has resulted in a surge of deepfakes that are present online and social media platforms.

The literature on deepfakes has provided clear definitions of the phenomenon. In the article “Deep Learning for Deepfakes Creation and Detection: A Survey,” the authors assert that deepfakes are a form of "artificial intelligence-synthesized content". The authors also describe a typology of deepfakes, which includes face swap, lip-sync, and

puppet master. Specifically, face swap deepfakes involve overlaying the face of a target person onto a video of a source person, resulting in a video of the target person appearing to do or say things they did not do. Lip-sync deepfakes manipulate videos so that lip movements match the audio, while puppet master deepfakes animate a person to follow the movements and expressions of another person's face, eyes, and head. This relationship is referred to as puppet and master. [20]

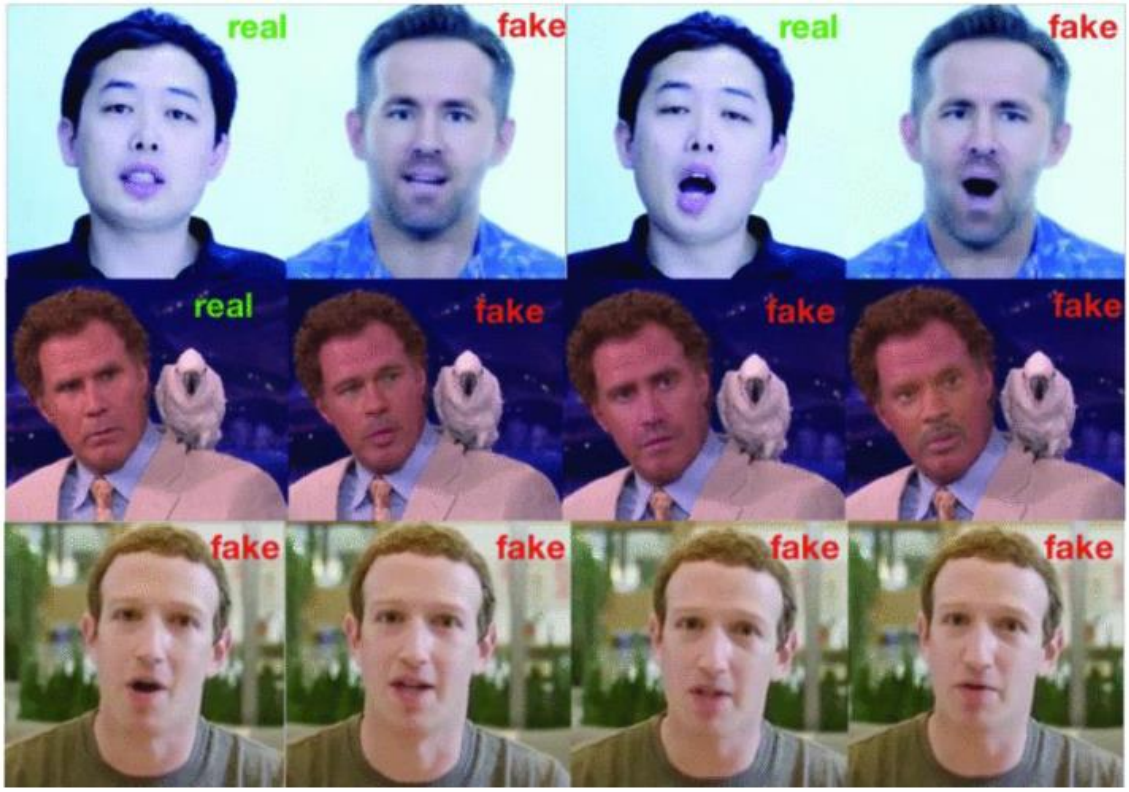

Figure 2 Example: Deepfake videos: (top) Head puppetry, (middle) face swapping, and (bottom) lip syncing [20]

An alternate definition of "Deepfakes" is presented, which includes a "broad scope of synthetic images, video, and audio generated through recent breakthroughs in the field of Machine Learning (ML), specifically in deep learning". This definition is cited from the report "The Deepfakes: A Grounded Threat Assessment" by CSET (Center for Security and Emerging Technology). The authors of the report emphasize that the term generally encompasses media that has been modified or generated by both ML and non-ML techniques. [21] The authors note that the deepfakes used in malicious disinformation campaigns are not sophisticated and lack quality, with the perpetrators focusing on the scale and mass production and spread of the content. According to the report, the most

common operational posture exhibited in online influence campaigns involves cheaply crafted audio-visual content and a lack of investment in high-quality deepfakes.

In the study "Deep Fakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence", Paris and Donovan present a typology of deepfakes based on the underlying technology used. [22] This spectrum of audio-visual manipulation encompasses both high-tech deepfakes created with cutting-edge, AI-reliant techniques, as well as "cheap fakes" produced with readily available and inexpensive software. Additionally, conventional methods such as speeding, slowing, cutting, re-staging, or re-contextualizing footage can also be used for manipulation purposes. [22]

The deepfake process relying on experimental machine learning represents the most computationally reliant and least publicly accessible means of creating deceptive media, whereas "cheap fakes" are the opposite end of the spectrum. Nevertheless, both types of manipulation can have significant impacts on politics and can manipulate evidence in various ways. [22]

The study by Paris and Donovan categorizes deepfakes based on the technology used, ranging from more technically sophisticated deepfakes created using AI-based methods to cheap fakes produced with easily accessible software. [22] The study lists FakeApp/After Effects, Video Dialogue Replacement (VDR) model, Generative Adversarial Networks (GANs), Recurrent Neural Network (RNN), Hidden Markov Models (HMM), and Long- and Short-Term Memory (LSTM) Models as the different types of deepfakes.

Belfer Centre's study classifies “The term “deepfake”—as hybrid of the terms “deep learning” and “fake”—first appearing on Reddit in 2017 and quickly becoming part of the technical lexicon” [23] based on the level of manipulation and content generation involved. The study categorizes deepfakes from "cheap fakes" to more technically sophisticated deepfakes. The "cheap fakes" category includes techniques such as re-contextualizing, lookalikes, speeding and slowing, face swapping-Rotoscope, lip-syncing, face replacement, synthetic speech production, face re-enactment and audio modulation, and face generation.

In both studies, the less manipulation and amendments to an existing audio-visual or visual content and more content generation with the use of ML, the more sophisticated is

the deepfake. The studies stress the importance of face swapping in deepfakes, and list various techniques and software used for this purpose.

2.2 Overview of Techniques for Analysing AI-Generated Content

In the next section of this paper, I cover two critical aspects of AI-generated content in cybersecurity, in detections and filtering techniques employ various methods to identify and prevent malicious content, while the challenges and limitations of existing methods include the difficulty in distinguishing between real and fake content, constantly evolving AI-generated attacks, and lack of standardization.

2.2.1 Detection and Filtering Techniques

Detection and filtering techniques for deepfakes have become increasingly crucial due to the rising prevalence of manipulated audio-visual content. These techniques utilize diverse methods to identify deepfakes, such as scrutinizing inconsistencies in facial features, audio, and video. One widely adopted approach entails employing machine learning algorithms to train models on a vast corpus of genuine and fake videos, enabling the system to detect patterns in the data and classify new videos as authentic or manipulated. Other techniques entail embedding invisible codes or signatures into the content using digital watermarking, which can be used to authenticate the content. However, prior research indicates that automatic detection tools have a short lifespan and may not be effective in the long term due to the continuous advancement of deepfake technology. [24] Consequently, researchers have been developing techniques continuously to tackle the challenges associated with deepfake detection, including the need for real-time detection and removal of deepfakes and the constant evolution of deepfake technology.

There are a variety of detection methods, and they vary from algorithms to forensic techniques, a couple are as follows. In the research paper titled "Deepfake Video Detection Using Recurrent Neural Networks," a novel system for detecting deepfake videos is proposed, which is based on a combination of various deep learning components. [25] The system consists of a Convolutional Neural Network (CNN) that extracts features from each frame of the video. The extracted features are then passed through a Long Short-Term Memory (LSTM) network for sequence processing. The LSTM network constructs a sequence of frames with features, which is then fed into the

Convolutional Long Short-Term Memory (Conv-LSTM) network. The Conv-LSTM network processes the input sequence and produces a sequence descriptor that computes the probabilities of the frame sequence being either a pristine or a deepfake video. [25]

In the research article titled "Protecting World Leaders Against Deep Fakes", the authors propose a forensic approach to identifying deepfakes by analysing facial behaviour. [26] The technique employs the use of OpenFace2 toolkit to extract distinctive facial and head movements from a video. The researchers contend that these movements are individualistic to each person, making it possible to detect inconsistencies and deviations in patterns when comparing the extracted expressions from deepfakes. In this manner, the expressions of impersonators can be easily distinguished from those of the original individual. Another method involves analysing inconsistencies in audio and video data. For instance, in the paper "An Overview of Recent Work in Media Forensics: Methods and Threats" [27], the authors propose analysing audio data for differences in background noise, reverberation, and other audio characteristics that may indicate manipulation. Similarly, video data can be analysed for inconsistencies in lighting, shadows, and other visual features that may not align with reality.

2.2.2 Challenges and Limitations of Existing Methods

Despite the advancements in deepfake generation and detection techniques, there are several challenges and limitations associated with existing methods. One major challenge is the rapid evolution of deepfake technology, with new algorithms and tools constantly emerging. This poses a significant challenge for detection methods that may struggle to keep pace with the evolving sophistication of deepfakes. As a result, deepfakes generated using state-of-the-art techniques may be more difficult to detect using existing methods, thereby limiting their effectiveness in identifying newly created deepfakes.

The generation of deepfakes is not without its limitations. One major limitation is the quality of the generated deepfakes. Despite the advancements in generative models, deepfakes may still exhibit noticeable artifacts, such as blurry edges, unrealistic lighting, or misaligned facial features, which can be detected by keen observers. These limitations in visual quality may reduce the believability and effectiveness of deepfakes, particularly in situations where scrutiny or detailed analysis is applied.

Another challenge is the potential for adversarial attacks on deepfake detection methods. Adversarial attacks refer to the intentional manipulation of deepfakes or their input data to evade detection. This can involve altering or adding imperceptible perturbations to the deepfake, making it difficult for detection methods to accurately classify them. Adversarial attacks can significantly reduce the effectiveness of existing detection methods, as they may fail to detect deepfakes that have been specifically manipulated to evade detection.

Additionally, existing methods for deepfake detection often rely on the availability of labeled training data, which may be scarce or impractical to obtain. Labeled data is crucial for training supervised machine learning models, but collecting and annotating a large and diverse dataset of deepfake videos can be time-consuming, expensive, and may raise ethical concerns related to the use of potentially harmful content. The lack of comprehensive and diverse labeled data can limit the accuracy and generalizability of deepfake detection methods, as they may not be able to effectively detect deepfakes that differ significantly from the training data. This can be a significant limitation in real-world scenarios where new and unknown types of deepfakes may emerge.

In conclusion, while existing methods for generation and detection of deepfakes have made considerable progress, they are not without challenges and limitations. The rapidly evolving nature of deepfake technology, the potential for adversarial attacks, and the reliance on labeled training data are some of the key challenges that may limit the effectiveness of existing methods. Addressing these challenges and developing robust and scalable solutions for deepfake generation and detection are critical to effectively combat the spread of deepfakes and their potential harms. Further research and innovation are needed to overcome these limitations and develop more robust methods for generating and detecting deepfakes in a constantly evolving technological landscape.

3 Research Methodology

The research methodology for analysing open-source deepfake techniques involves a comprehensive and systematic approach to examining the technical aspects of deepfakes. The research begins with a look into existing theoretical background of the relevant studies and a literature review had been conducted by the author on deepfakes and open-source techniques. This review will help identify the most commonly used open-source deepfake techniques and the potential risks and benefits of using these techniques. The next step involves selecting a representative sample of open-source deepfake techniques and analysing them using various metrics such as accuracy, complexity, and computational requirements. The research will also involve generating and attempting to detect generated content in an effort to examine potential impact of the software. Finally, the research will conclude with a discussion of the implications of the findings for future research on deepfakes and their impact on society.

3.1 Overview of Proposed Method

The proposed method involves evaluating the effectiveness of two open-source deepfake generation tools and one deepfake detection tool, as well as one website detection tool keeping in mind their accessibility to the average individual. The deepfake generation tools to be evaluated are selected based on their popularity, ease of use, and computational and storage requirements. The generated deepfakes will be run through the deepfake detection tool to assess its ability to accurately identify them. Additionally, a website detection tool will be used to evaluate its effectiveness in identifying deepfakes that have been uploaded to a website. The proposed method aims to provide a comprehensive analysis of the deepfake generation and detection tools, as well as website detection tools, to better understand their strengths and limitations in detecting and preventing the spread of malicious deepfakes. The results of this analysis could inform the development of more effective detection and filtering techniques to mitigate the potential harms associated with deepfakes.

3.2 Data Collection and Pre-processing

The collection and pre-processing of data are critical components in the development of AI models. In this thesis, we focus on the data collection and pre-processing steps involved in developing a deepfake model from two videos featuring an individual each.

Once the dataset is collected, pre-processing is performed to prepare the data for training. This includes standardizing the resolution and colour space of the images or videos, as well as applying data augmentation techniques if needed to increase variability and diversity. Bias in the dataset must be addressed to ensure the resulting model is not perpetuated with biases related to gender, race, or age.

Data cleaning is another critical part of pre-processing. This involves the removal of low-quality images or videos, correcting misalignments, and filtering out irrelevant or redundant data. Thorough data cleaning is necessary to maximize data quality and eliminate any noise or artifacts that could negatively impact the deepfake model's quality.

Overall, the quality, accuracy, and ethical implications of the resulting deepfake models are directly influenced by the data collection and pre-processing steps.

3.3 Evaluation Metrics

The evaluation metrics are ones conserved about practical aspects such as ease of use, time efficiency, and computational requirements when evaluating generation and detection tools for deepfakes. Ease of use is a crucial factor as it affects the practicality and adoption of these tools by end-users. User-friendly interfaces, clear documentation, and availability of tutorials or guides can greatly impact the usability of the tools. The time it takes to generate or detect deepfakes is another important consideration, as real-time or near-real-time performance may be required in certain applications, such as content moderation or live video analysis. The computational requirements, including hardware and software dependencies, processing power, and memory usage, can significantly affect the practicality and scalability of these tools.

Ease of use: This will be assessed through user feedback, surveys, or usability testing to evaluate how user-friendly and intuitive the tool's interface and workflow are, including the availability of documentation, tutorials, and support.

Computational requirements: This will be assessed by evaluating the hardware and software dependencies of the tool, such as the minimum and recommended system specifications, processing power, memory usage, and compatibility with different operating systems or hardware configurations.

Storage requirements: This metric measures the amount of storage space required for storing generated deepfakes or training data, which can be relevant in scenarios where storage space is a constraint, such as in resource-limited environments or when handling large datasets.

Time efficiency: This will be measured by evaluating the time it takes for the generation or detection tool to process a given input, such as the average time taken per image or video, or the time taken for real-time or near-real-time analysis.

4 Experiments and Results

For our evaluation, we used two videos each of a different person created after following the gathering of video footage is specific style and extracting features from the face region using tools. We then compared the performance of different deepfake generation and detection software on videos after altering then with DeepFaceLab and FaceSwap, chosen in particular because of meeting the criteria of being open-source, in continuous development, having substantial documentation, being compatible with modern systems, and having transparency with the algorithms and methodology used.

4.1 Software used/Model Selection and Training

Figure 3 DeepFacelab cmd overview

DeepFaceLab (DFL) offers a flexible workflow that encompasses three distinct phases, namely extraction, training, and conversion. [28] These phases are sequentially presented in DFL, where the pipeline is abstracted into these three key components. It is worth mentioning that DFL adheres to a typical one-to-one face-swapping paradigm, where the data is categorized into two types: "src" and "dst," which stand for source and destination, respectively. These terminologies will be used consistently throughout the later narrative.

The pipeline for DeepFaceLab typically involves several stages, including data collection, pre-processing, model training, and post-processing. Here is a detailed overview of the typical DeepFaceLab pipeline:

1. Data Collection: The first step in the DeepFaceLab pipeline is to collect the data required for training the deepfake generation model. This typically involves gathering a large dataset of real images and videos of the target person whose face will be swapped in the deepfake. The dataset may include various poses, expressions, and lighting conditions to ensure diversity and generalization of the model.
2. Pre-processing: Once the data is collected, pre-processing is performed to prepare the data for training. This may involve resizing, cropping, and aligning the images or videos to ensure that the faces are centered and have consistent orientations. Pre-processing may also involve data augmentation techniques, such as flipping, rotating, or changing brightness/contrast, to increase the diversity and robustness of the training data.
3. Model Training: After pre-processing, the next step is to train the deepfake generation model. DeepFaceLab supports different deep learning architectures, such as Convolutional Neural Networks (CNNs) or Generative Adversarial Networks (GANs), which are trained using the collected and pre-processed data. The training process involves optimizing the model parameters to minimize the discrepancy between the real and generated images or videos. This typically requires several iterations of training, validation, and fine-tuning to achieve desired results.
4. Post-processing: Once the model is trained, post-processing is applied to refine the generated deepfake images or videos. This may involve techniques such as color correction, tone matching, blending, and texture smoothing to make the deepfake more visually convincing and realistic. Post-processing may also involve adding additional effects, such as noise, blur, or sharpening, to further enhance the visual quality of the deepfake.
5. Detection: In addition to generation, DeepFaceLab also includes features for deepfake detection. This typically involves training a separate detection model using a labeled dataset of real and fake images or videos. The detection model is

then used to classify new images or videos as real or fake based on their features, such as facial landmarks, textures, and artifacts.

6. Evaluation: Finally, the generated deepfake images or videos are evaluated using appropriate metrics, as discussed in the previous responses, to assess their quality, authenticity, and performance. This may involve subjective evaluations by human raters or objective evaluations using established image or video quality metrics.

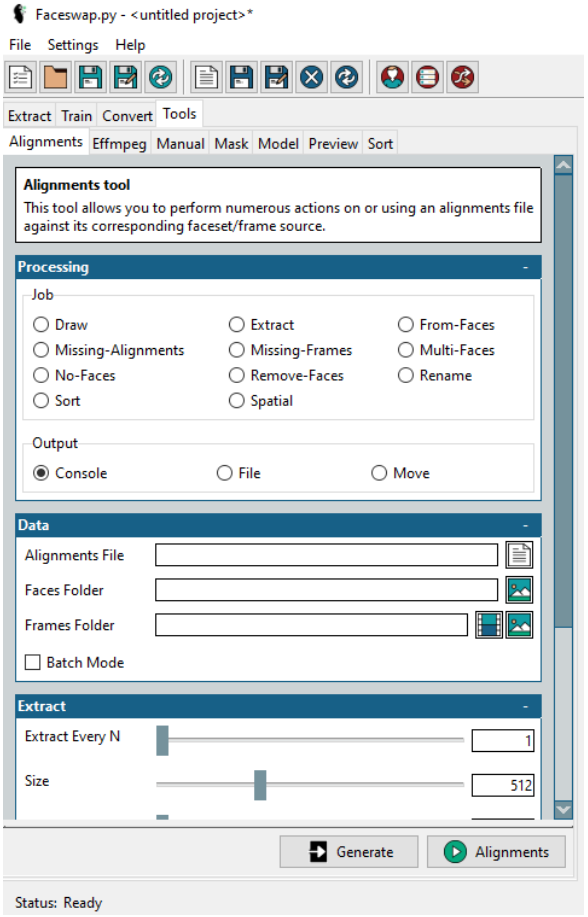

Figure 4 FaceSwap GUI

FaceSwap is another popular open-source deepfake generation tool that provides a graphical user interface (GUI) for ease of use. [29] While it shares similarities with DeepFaceLab in terms of the general pipeline, there are some differences in the specific steps and workflow. Here is a detailed overview of the FaceSwap pipeline, highlighting the differences compared to DeepFaceLab:

1. Data Collection: Similar to DeepFaceLab, the first step in the FaceSwap pipeline is to collect the data required for training the deepfake generation model. This typically involves gathering a large dataset of real images and videos of the target person and the source person whose face will be swapped in the deepfake. The dataset may include various poses, expressions, and lighting conditions to ensure diversity and generalization of the model.
2. Pre-processing: Once the data is collected, pre-processing is performed to prepare the data for training. This may involve resizing, cropping, and aligning the images or videos, similar to DeepFaceLab. However, FaceSwap also provides a GUI-based interface for manual facial landmark annotation, which allows the user to define the facial landmarks more accurately.
3. Model Training: After pre-processing, the next step is to train the deepfake generation model. FaceSwap uses a deep learning architecture based on autoencoders, which are trained using the collected and pre-processed data. The training process involves optimizing the model parameters to minimize the discrepancy between the real and generated images or videos, similar to DeepFaceLab.
4. Post-processing: Once the model is trained, FaceSwap provides a GUI-based interface for post-processing to refine the generated deepfake images or videos. This includes features such as color correction, tone matching, blending, and texture smoothing, similar to DeepFaceLab. However, FaceSwap's GUI-based interface allows for more intuitive and interactive post-processing adjustments, making it user-friendly for users without extensive technical expertise.
5. Detection: FaceSwap does not include built-in deepfake detection features like DeepFaceLab. However, users can use external deepfake detection tools or libraries to detect the generated deepfake images or videos.
6. Evaluation: Finally, the generated deepfake images or videos can be evaluated using appropriate metrics, as discussed in the previous responses, to assess their quality, authenticity, and performance. This may involve subjective evaluations by human raters or objective evaluations using established image or video quality metrics, similar to DeepFaceLab.

4.2 Overview of the Dataset and Features

The dataset used for this experimentation done was using the faces of two consenting individuals. Two videos following the following steps were taken in order to be used as bases for the deepfake processing.

The following detailed steps for the facial expressions and movements performed in front of the camera were given to the volunteers to capture enough visual data for training a deepfake model:

1. Neutral Expression: Start with a relaxed and neutral expression, looking straight into the camera. Keep your face relaxed and natural, with a neutral mouth position, and avoid any exaggerated movements.
2. Smiling: Gradually smile, starting with a small smile and gradually increasing to a bigger smile. Try different variations of smiles, such as closed-lip smiles, teeth showing smiles, and smiles with raised cheeks.
3. Frowning: Create a frown expression by lowering your eyebrows, puckering your lips, and pulling the corners of your mouth downward. Experiment with different intensities of frowns to capture a range of expressions.
4. Surprise: Open your eyes wide and raise your eyebrows to create a surprised expression. Keep your mouth slightly open to complete the surprised look. Try different variations of surprise, from mild to intense.
5. Anger: Create an angry expression by furrowing your eyebrows, narrowing your eyes, and clenching your jaw. Experiment with different intensities of anger, from subtle to intense.
6. Blinking: Blink your eyes naturally as you would in real life. Blinking adds realism to the deepfake model and helps in capturing the natural eye movements.
7. Head Movements: Move your head slowly in different directions, such as tilting it to the left or right, nodding up and down, or shaking it side to side. This helps in capturing the natural movements of the head and neck.
8. Speech Movements: Say different words, phrases, and sentences with varying intonations and mouth movements. This helps in capturing the movement of your lips, tongue, and jaw during speech, which is essential for realistic lip synchronization in the deepfake model.

9. Emote with Expressions: Experiment with different emotions like happiness, sadness, surprise, anger, and disgust, as well as subtle micro-expressions that convey emotions. This helps in capturing a wide range of facial movements and expressions for the deepfake model.
10. Relax and Reset: Take breaks in between expressions and relax your face to reset it to a neutral expression. This helps in avoiding muscle strain and capturing a clean reference for the deepfake model.

These expressions and movements were attempted to be performed naturally and comfortably, as unnatural, or exaggerated movements may result in less realistic deepfake results. A length of 5 minutes was considered enough footage with sufficient variations of facial expressions and movements to provide ample visual data for the deepfake model to learn from.

Afterwards these videos were run through the respective extractors of DeepFaceLab and Faceswap to generate their datasets for model creation and usage.

Figure 5 Model (DeepFake) training with FaceSwap

4.3 Performance Evaluation and Comparison of the Results

Edited Video with software and model	Mesonet Detection (Deepstar)	Avatarify	Deepware	Seferbekov	Ensemble
Face A to B (DeepFaceLab)	True	Unsure	Detected	Detected	Detected
Face B to A (DeepFaceLab)	True	Unsure	Detected	Detected	Detected
Face A to B (FaceSwap)	True	Detected	Detected	Not Detected	Detected
Face B to A (FaceSwap)	True	Detected	Detected	Detected	Detected
High Quality Fake (DeepFaceLab)	False	Unsure	Not Detected	Not Detected	Not Detected
High Quality Fake (FaceSwap)	False	Unsure	Not Detected	Detected	Detected

Table 1 Overview of Detection of Generated Content

The software was run on the following hardware and software CPU Intel Core i7 9750H @ 2.60GHz, RAM 16.0GB Dual-Channel DDR4, Graphics card 4095MB NVIDIA GeForce GTX 1650, Storage 1TB SSD. Even with the mentioned hardware the process involved in generating a single 3-minute video using two faces of rather similar appearance took about 16 hours to generate with 160000 thousand iterations made during the modelling process using FaceSwap as a lesser example.

DeepFaceLab is a popular open-source software used for creating deepfake videos. In terms of ease of use, the software requires a basic understanding of computer

programming and image processing. While the interface is user-friendly, the process of training models can be complex and time-consuming for beginners. In terms of computational requirements, the software requires high-end graphics processing units (GPUs) to train deep learning models effectively. This can be a limiting factor for users with low-end hardware. The storage requirements are also high due to the large size of the training data and the models themselves. The time efficiency of DeepFaceLab depends on the complexity of the project and the hardware used. For complex projects, the training time can take several days or even weeks. In summary, DeepFaceLab is a powerful tool for creating high-quality deepfake videos, but it requires some technical expertise and high-end hardware to use effectively.

FaceSwap is an open-source deepfake software that allows users to create realistic face swaps in images and videos. In terms of ease of use, FaceSwap has a user-friendly interface that is easy to navigate, and the software supplies detailed documentation and tutorials to help beginners get started. The computational requirements for FaceSwap are moderate, and the software can run on standard CPUs or GPUs. The storage requirements are also moderate, as the software uses image and video files as input data. Time efficiency is one of the strengths of FaceSwap, as the software uses a pre-trained model to speed up the face swapping process. The time required for face swapping depends on the complexity of the project and the hardware used, but in general, FaceSwap is considered to be faster than other deepfake software. In summary, FaceSwap is a user-friendly and efficient deepfake software that can run on standard hardware, making it accessible to a wide range of users.

DeepStar is a deepfake detection tool that uses deep learning algorithms to detect the presence of manipulated media. The software is designed to analyse the underlying features of an image or video to identify whether it has been altered or manipulated. DeepStar works by comparing the features of the original image or video with the features of the suspected manipulated media. The software can detect several types of deepfake techniques, including face swaps, image blending, and object removal. One of the strengths of DeepStar is its accuracy, as the deep learning algorithms used in the software are highly effective at identifying manipulated media. The tool is also user-friendly and can be used by individuals or organizations to detect deepfakes in a range of contexts, such as social media, news articles, and political propaganda. Overall, DeepStar is a

powerful and effective deepfake detection tool that can help combat the spread of manipulated media and disinformation. [30]

Mesonet detection with DeepStar refers to the use of a multi-model ensemble approach to deepfake detection. The Mesonet approach involves combining the outputs of multiple deep learning models to improve the overall accuracy and robustness of deepfake detection. In the context of DeepStar, Mesonet detection involves using a combination of deep learning models to analyse the features of an image or video and detect the presence of manipulated media. Each model in the Mesonet has different strengths and weaknesses, and by combining them, the overall accuracy of deepfake detection is improved. The use of Mesonet detection with DeepStar has been shown to be highly effective in detecting a wide range of deepfake techniques, including face swaps, image blending, and object removal. [30]

Scanner.deepware.ai is a web-based deepfake detection tool provided by Deepware, a company that specializes in AI-driven media forensics. The website allows users to upload image or video files for analysis and provides a confidence score indicating the likelihood of the media being a deepfake. The tool uses machine learning algorithms to analyse various features of the media, such as facial expressions and movements, to determine if the content has been manipulated or generated using AI. It also has the ability to detect deepfakes that have been compressed or re-encoded to hide manipulation. Overall, scanner.deepware.ai provides an easy-to-use and accessible option for individuals or organizations to detect deepfakes in their media. [31]

The Deepware scanner is a proprietary deepfake detection model developed by Deepware, which is designed to detect deepfake videos with high accuracy. It uses a combination of computer vision techniques and machine learning algorithms to analyze a video frame by frame and identify any signs of manipulation or synthetic generation.

The Seferbekov model is another deepfake detection algorithm that was developed by a team of researchers at the Moscow Institute of Physics and Technology. It uses a machine learning approach to analyze facial expressions, movements, and other features of a video to determine whether it is a deepfake or not.

The Ensemble model is a combination of multiple deepfake detection models with Seferbekov and Deepware being the main ones, which have been trained on different

datasets and use different techniques to identify deepfakes. By combining the results of multiple models, the Ensemble model can provide more accurate and reliable results compared to any single model.

4.4 Summary of Experiment

In this thesis experiment, we aimed to evaluate the performance of two open-source deepfake generation tools, DeepFaceLab and FaceSwap, based on several criteria, including ease of use, computational requirements, storage requirements, and time efficiency. We conducted a detailed review of each tool based on these criteria, considering their user interfaces, hardware requirements, and training times. Following this review, we generated a couple of deepfakes using both tools and ran them through several deepfake detection tools, including DeepStar and other models. We then tabulated the results and analysed them to determine the effectiveness of the tools.

Our findings suggest that while DeepFaceLab and FaceSwap can be effective in generating high-quality deepfakes, they come with some caveats. In terms of ease of use, Faceswap had the more user-friendly interface, but the training process can be complex and time-consuming for beginners concerning both. The computational and storage requirements of the tools are moderate, with DeepFaceLab requiring higher-end hardware for training. Time efficiency varies depending on the complexity of the project, but in general, DeepFaceLab is slower than FaceSwap due to its more complex training process.

Regarding the deepfake detection results, we found that while the generated deepfakes can at times look convincing to the human eye, they were easily detected by deepfake detection tools such as DeepStar and other models. However, this mainly true for the ones done on a relatively fast basis. If the detection tools were given an above average example of a generated deepfake the tools at time do fail to detect the change. The experiment also revealed that the effectiveness of the deepfakes generated using these tools is highly dependent on the quality of the input data and the skill of the user. This indicates that while the tools can produce convincing deepfakes, they can be detected if not enough time or effort is spent on the generation process. Our analysis also showed that the effectiveness of the detection tools varied depending on the type of deepfake generated, with some tools performing better than others.

Another limitation of this experiment is to keep in mind that when considering the reliability of the detections using the given detection tools and algorithms. False positives and negatives are possible, and overall reliability of detection cannot be assured but speculated by considering overall performances of the whole detection process. It is important to note that the reliability of detection tools and algorithms can be influenced by several factors, such as the quality of the training data used to develop the algorithms, the complexity of the deepfake content being analysed, and the sophistication of the detection methods being employed. Additionally, the rapidly evolving nature of deepfake technology means that detection methods may quickly become outdated and ineffective.

In conclusion, our experiment highlights the potential capability of using open-source deepfake generation tools such as DeepFaceLab and FaceSwap. While these tools can be effective in generating high-quality deepfakes, they come with some limitations and can be detected by current deepfake detection methods. However, the detection tools are not fool proof and can be fooled given more effort was spent on the generation period. Therefore, it is essential to approach the use of these tools with caution and use further means of protection and guarantee for safety. Furthermore, more research is needed to improve the detection methods and to develop robust countermeasures against the malicious use of deepfakes.

5 Conclusion

With the widespread availability of social media platforms, a vast amount of content is publicly accessible, providing a large pool of potential targets for deepfake creation. This holds true for individuals belonging to diverse backgrounds, living in cities, villages, or any social group. Moreover, creating deepfakes does not need significant investment, and it is not limited to national states or corporations, as individuals and small criminal groups can also initiate deepfake attacks. The goal of this paper was to examine how effective the generation and detection of open-source tools were regarding this topic, and what an individual or organization could do with malicious intentions and what they could do to protect themselves from such attempts.

The findings of this study suggest that individuals with immediate access to the internet and the tools at hand can create edited content using various AI tools available. However, the quality of the output is limited by the amount of time taken to gather resource footage of sufficient quality and quantity, the time taken learn a new tool and the system requirements of the tool. Both DeepFaceLab and FaceSwap as the representative tools taken in the paper were found to be effective in generating deepfakes, but they come with some caveats. FaceSwap was easier to use, had moderate computational and storage requirements, and was relatively efficient in terms of time. On the other hand, DeepFaceLab was more complex to set up, required high-end resources, and had longer training times, but on average generated higher quality media.

In terms of deepfake detection, open-source tools were available but limited in quantity and usability. The results of detection were solely dependent on the capability of individual tools and reliability of detection of fake media was in question due to possible false positives and negatives and failure to detect fakes. The results showed that while the detection tools were generally effective, they could still be circumvented if sufficient time and effort were put into the generation process of edited media. The detection tools used in this study were able to identify various artifacts in the deepfakes, such as misaligned features, inconsistent lighting, and unnatural movements. However, as the sophistication and complexity of deepfakes continue to increase, it is crucial for researchers to continue developing more sophisticated and reliable detection algorithms to stay ahead of the threat of deepfake attacks.

The experiment's summary shows that on one hand the currently available tools are capable of creating highly convincing deepfakes. On the other hand, relying solely on detection tools as a guarantee of protection against malicious parties intending to use deepfakes is not advisable.

To mitigate the impact of deepfake attacks, individuals and organizations could implement several measures. For authentication of sensitive or critical accounts, a multi-factor authentication approach should be standard. Authentication could also be done including three basic factors: something that the user has, something that the user knows, and something that the user is. Carefully choosing the "something" items can render deepfakes ineffective. Furthermore, personnel awareness training, is necessary for financial organizations to identify red flags associated with deepfake technology. Individual could minimize the exposure of high-quality personal images on social media, and for verification of sensitive accounts, prioritize the use of biometric patterns that are less exposed to the public, such as irises and fingerprints. Finally, significant policy changes are required to address the issue at a larger scale, including the use of current and previously exposed biometric data and preparation for the future state of cybercriminal activities.

In conclusion, the study reveals the potential dangers posed by deepfake technology and emphasizes the need for ongoing research and development of advanced detection algorithms. Although open-source AI tools have proven effective in generating realistic fake content, the current detection methods have their limitations, and cybercriminals are constantly evolving their tactics. Therefore, it is essential to remain vigilant and prepare for the future by implementing effective policies and countermeasures to safeguard against malicious use of deepfakes.

6 Future Work

Potential future work in this field can focus on the development of countermeasures to deepfakes. One promising approach is the use of blockchain technology to create a tamper-proof digital ledger that can be used to verify the authenticity of content. Additionally, there is a need to continue developing more effective deepfake detection algorithms to keep up with the evolving capabilities of deepfake generation. Furthermore, research could explore the use of machine learning techniques to improve the accuracy of deepfake detection and reduce false positives. Finally, it would be valuable to conduct studies to investigate the effectiveness of various anti-deepfake measures, such as watermarking and digital signatures, in mitigating the potential harms of deepfakes. The development of such measures will be crucial in the ongoing battle against the malicious use of deepfake technology.

References

[1] Thomas Brewster, “Fraudsters Cloned Company Director’s Voice In $35 Million Bank Heist, Police Find”, Oct 14 2021. [Online]. Available: https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=6dec67f97559

[2] John Letting, “How to tell reality from a deepfake?” April 1 2021 [Online] Available: https://www.weforum.org/agenda/2021/04/are-we-at-a-tipping-point-on-the-use-of-deepfakes/

[3] “VMware Report Warns of Deepfake Attacks and Cyber Extortion” August 8, 2022 [Online] Available: https://news.vmware.com/releases/vmware-report-warns-of-deepfake-attacks-and-cyber-extortion

[4] Bolshunov, P., Marcel, S. “Deepfakes: A new threat to face recognition? Assessment and Detection”, 20 Dec 2018, [Online] Available: https://arxiv.org/pdf/1812.08685.pdf.

[5] R. Chesney and D. K. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," December 2019, [Online] Available: https://scholarship.law.bu.edu/faculty_scholarship/640/

[6] Delfino, R.A. “Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act”, December 09, 2019, [Online] Available: https://ir.lawnet.fordham.edu/flr/vol88/iss3/2/

[7] Feldstein, S. “How Artificial Intelligence Systems Could Threaten Democracy”, April 24 2019, [Online] Available: https://carnegieendowment.org/2019/04/24/how-artificial-intelligence-systems-could-threaten-democracy-pub-78984

[8] Alec Banks, “Op-Ed | Deepfakes & Why the Future of Porn is Terrifying” Highsnobiety. 2020, Available: https://www.highsnobiety.com/p/what-are-deepfakes-ai-porn/.

[9] Kevin Roose. “Here come the fake videos, too”. The New York Times, 4 March 2018, [Online], Available: https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html

[10] K. Quach, "Politically linked deepfake LinkedIn profile sparks spy fears, Apple cooks up AI transfer tech, and more," 17 June 2019. [Online]. Available: https://www.theregister.com/2019/06/17/roundup_ai/ .

[11] The Conversation, "3.2 billion images and 720,000 hours of video are shared online daily. Can you sort real from fake?," November 3, 2020, [Online]. Available: https://theconversation.com/3-2-billion-images-and-720-000-hours-of-video-are-shared-online-daily-can-you-sort-real-from-fake-148630.

[12] Borges-Tiago, T.; Tiago, F.; Silva, O.; Guaita Martinez, J.M.; Botella-Carrubi, D. “Online users’ attitudes toward fake news: Implications for brand management.” September 2020 [Online] Available: https://www.researchgate.net/publication/339903984_Online_users%27_attitudes_toward_fake_news_Implications_for_brand_management

[13] Figueira, Á.; Oliveira, L. “The current state of fake news: Challenges and opportunities.” Procedia Comput. Sci. 2017, [Online] Available: https://www.sciencedirect.com/science/article/pii/S1877050917323086
[14] Anderson, K.E. “Getting Acquainted with Social Networks and Apps: Combating Fake News on Social Media”; Library Hi Tech News; Emerald Group Publishing Limited: Bingley, UK, 2018.[Online] Available: https://scholarship.libraries.rutgers.edu/esploro/outputs/acceptedManuscript/Getting-acquainted-with-social-networks-and/991031550143704646
[15] Zannettou, S.; Sirivianos, M.; Blackburn, J.; Kourtellis, N. “The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans." 10 Apr 2018, [Online] Available: https://arxiv.org/pdf/1804.03461.pdf
[16] Borges, P.M.; Gambarato, R.R. “The role of beliefs and behavior on facebook: A semiotic approach to algorithms, fake news, and transmedia journalism.” Int. J. Commun. 2019, 13, 16 [Online] Available: https://ijoc.org/index.php/ijoc/article/view/10304/2550
[17] Vladimir Kropotov, Fyodor Yarochkin, Craig Gibson, Stephen Hilt, “How Undergound Groups Use Stolen Identities and Deepfakes”, 27 September 2022, [Online] Available: https://www.trendmicro.com/en_us/research/22/i/how-underground-groups-use-stolen-identities-and-deepfakes.html
[18] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. 25 Jan 2019, [Online] Available: https://arxiv.org/pdf/1901.08971.pdf
[19] Y. Li, M.-C. Chang and S. Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking,", 11 June 2018. [Online]. Available: https://arxiv.org/pdf/1806.02877.pdf.
[20] T. T. Nguyen, Q. V. Hung Nguyen, C. M. Nguyen, D. Nguyen, D. T. Nguyen and S. Nahavand, “Deep Learning for Deepfakes Creation and Detection: A Survey,” 26 April 2021, [Online]. Available: https://arxiv.org/pdf/1909.11573.pdf.
[21] T. Hwang, “The Deepfakes: A Grounded Threat Assessment,” July 2020. [Online]. Available: https://cset.georgetown.edu/wp-content/uploads/CSET-Deepfakes-Report.pdf.
[22] J. Donovan and B. Paris, "Deep Fakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence," [Online]. Available: https://datasociety.net/wp-content/uploads/2019/09/DataSociety_Deepfakes_Cheap_Fakes.pdf.
[23] Belfer Center for Science and International Affairs, "TECH POLICY FACTSHEET: DEEPFAKES," 2020. [Online]. Available: https://www.belfercenter.org/sites/default/files/2020-10/tappfactsheets/Deepfakes.pdf.
[24] A. Engler, "Fighting deepfakes when detection fails," 14 November 2019. [Online]. Available: https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails.
[25] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," November 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8639163
[26] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano and H. Li, "Protecting World Leaders Against Deep Fakes," 2020. [Online]. Available: http://www.hao-li.com/publications/papers/cvpr2019workshopsPWLADF.pdf.
[27] Kratika Bhagtani, Amit Kumar Singh Yadav, Emily R. Bartusiak, Ziyue Xiang, Ruiting Shao, Sriram Baireddy, and Edward J. Delp “An Overview of Recent Work in Media

Forensics: Methods and Threats", 26 Apr 2022, [Online]. Available: https://arxiv.org/pdf/2204.12067.pdf
[28] Iperov, "DeepFaceLab", Dec 31, 2022 [Online] Available: https://github.com/iperov/DeepFaceLab
[29] Deepfakes, "FaceSwap" Feb 24 2023 [Online] Available: https://github.com/deepfakes/faceswap
[30] PenTestIT, "Deepstar: An Open-source Deepfake Detection Toolkit" 2020, [Online] Available: https://pentestit.com/deepstar-open-source-deepfake-detection-toolkit/
[31] Deepware, "Face detection and recognition library that focuses on speed and ease of use" Feb 8, 2021, [Online] Available: https://github.com/deepware/dface

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Rashad Gafarli

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis „Analysis of AI-generated Content: Open-Source Techniques and Countermeasures “, supervised by Fuad Budagov.
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

15.05.2023

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.