EMERGING TECHGAURD
ON DEEP FAKE DETECTION

A PROJECT REPORT

Submitted by

Anshika Mittal (20BCS9408)

Arzit Mahajan (20BCS7736)

Gurleen Kaur (20BCS7403)

Ms. Ishu Sharma (20BCS7737)

Vigya Vigy (20BCS9490)

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING IN

COMPUTER SCIENCE & ENGINEERING

Chandigarh University

NOV 2023

BONAFIDE CERTIFICATE

Certified that this project report **EMERGING TECHGAURD ON DEEP FAKE DETECTION** is the Bonafide work of **Anshika Mittal (20BCS9408), Arzit Mahajan (20BCS7736), Gurleen Kaur (20BCS7403), Ms. Ishu Sharma (20BCS7737), Vigya Vigy (20BCS9490)** who carried out the project work under our supervision.

SIGNATURE	**SIGNATURE**
Er. Adil Husain Rather	Ms. Navpreet Kaur Ahluwalia
SUPERVISOR	**HEAD OF THE**
ASST. PROFESSOR	**DEPARTMENT**
Computer science and Engineering	Computer science and Engineering

Submitted for the project viva-voce examination held on ______________________

INTERNAL EXAMINER	**EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

We are very grateful to all the people who have contributed in one way or the other to enable us to come up with this project. We wish to express our sincere and heartfelt gratitude to our supervisor Er. Adil Husain Rather for their guidance and support by going through this project, making recommendations and been available for consultation. We also greatly thank our family members for according to us moral support and encouragement during the project. Special thanks goes to the head of the department, Ms. Navpreet Kaur Ahluwalia and Department of Computer Science and Engineering for their distinctive professional guidance. I also sincerely thank you for the time spent proofreading and correcting my mistakes.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

The project "**EMERGING TECHGAURD ON DEEP FAKE DETECTION**" aims to allow people to be able to identify the fake images, audio, video etc. using deep learning models. The recent trends of emerging AI have made it easier for forgers to play with images and other media using deep fake. Deep fake technology uses artificial intelligence and machine learning to create highly convincing fake videos and audio recordings, often with malicious intent. To counter this threat, various methods and technologies have emerged to detect and mitigate deep fakes.

With the improvement of face and voice recognition systems, it will prove helpful to identify the discrepancies between original and fake media.

As deep fake technology advances, so do the techniques to detect them like GAN's (Generative Adversarial Networks). Other deep learning approaches are being developed to counteract GAN-generated deep fakes.

Advanced machine learning algorithms, including deep neural networks, are being developed to identify anomalies in videos or audio that may indicate the presence of deep fakes. These algorithms analyze patterns, inconsistencies, and artifacts that are not present in genuine media.

संक्षेप

(हिंदी)

प्रोजेक्ट "इमर्जिंग टेकगार्ड ऑन डीप फेक डिटेक्शन" का उद्देश्य लोगों को डीप लर्निंग मॉडल का उपयोग करके नकली छवियों, ऑडियो, वीडियो आदि की पहचान करने में सक्षम बनाना है। उभरते एआई के हालिया रुझानों ने जालसाजों के लिए डीप फेक का उपयोग करके छवियों और अन्य मीडिया के साथ खेलना आसान बना दिया है। डीप फेक तकनीक अक्सर दुर्भावनापूर्ण इरादे से अत्यधिक विश्वसनीय नकली वीडियो और ऑडियो रिकॉर्डिंग बनाने के लिए कृत्रिम बुद्धिमत्ता और मशीन लर्निंग का उपयोग करती है। इस खतरे का मुकाबला करने के लिए, गहरी नकली वस्तुओं का पता लगाने और उन्हें कम करने के लिए विभिन्न तरीके और प्रौद्योगिकियां सामने आई हैं।

चेहरे और आवाज पहचान प्रणालियों में सुधार के साथ, यह मूल और नकली मीडिया के बीच विसंगतियों की पहचान करने में मददगार साबित होगा।

जैसे-जैसे गहरी नकली तकनीक विकसित होती है, वैसे-वैसे उनका पता लगाने की तकनीक जैसे GAN (जेनरेटिव एडवरसैरियल नेटवर्क) भी विकसित होती है। GAN-जनित डीप फेक का मुकाबला करने के लिए अन्य गहन शिक्षण दृष्टिकोण विकसित किए जा रहे हैं।

वीडियो या ऑडियो में विसंगतियों की पहचान करने के लिए डीप न्यूरल नेटवर्क सहित उन्नत मशीन लर्निंग एल्गोरिदम विकसित किए जा रहे हैं जो डीप फेक की उपस्थिति का संकेत दे सकते हैं। ये एल्गोरिदम उन पैटर्न, विसंगतियों और कलाकृतियों का विश्लेषण करते हैं जो वास्तविक मीडिया में मौजूद नहीं हैं।

CHAPTER 1

INTRODUCTION

1.1 CLIENT IDENTIFICATION/NEED IDENTIFICATION

There are several contemporary issues and challenges associated with deep fake detection systems. It's important to note that the field of deep fake technology and detection is rapidly evolving, and new challenges may have emerged since then.

- Deep fake creation techniques are advancing rapidly, making it challenging for detection systems to keep pace. Newer deep fakes may be more convincing and harder to detect.

- Malicious actors are developing techniques to create deep fakes that are specifically designed to evade detection systems. Adversarial attacks can be used to fool detection algorithms.

- Deep fake detection often requires analyzing personal data (e.g., facial features, voice recordings). Balancing the need for detection with data privacy concerns is an ongoing challenge.

- Deep fakes created from limited data (e.g., a few photos or voice samples) can be challenging to detect because there is less genuine data for comparison.

- The evolution of generative models, such as Generative Adversarial Networks (GANs), contributes to the complexity of deep fake creation. As these models become more sophisticated, they enable the generation of content that closely mimics real footage, making it even more difficult for detection systems to distinguish between genuine and manipulated content.

- The integration of multiple modalities, such as combining altered videos with synthetic audio, adds another layer of intricacy. Detection systems that primarily focus on a single modality may struggle to identify deep fakes that leverage a combination of visual and auditory elements.

As deep fake technology advances, there is an increasing demand for real-time detection systems, especially in applications like social media platforms and live video streaming. Developing detection mechanisms that can operate swiftly without compromising accuracy poses a significant challenge.

- The absence of standardized benchmarks for evaluating deep fake detection systems complicates the assessment of their effectiveness. Varying datasets, evaluation metrics, and testing conditions make it challenging to compare the performance of different detection algorithms accurately.

- Determining the origin of a deep fake, or attributing it to a specific individual or group, is a persistent challenge. Malicious actors may intentionally obfuscate the source of the manipulated content, hindering efforts to hold responsible parties accountable.

- The use of deep fake detection systems raises ethical concerns, particularly regarding privacy and consent. Striking a balance between protecting individuals from the harmful effects of deep fakes and respecting their privacy rights is an ongoing ethical dilemma.

- Deep fake detection models trained on specific datasets may struggle to generalize effectively to new and diverse deep fake scenarios. Ensuring that detection systems remain robust across various contexts and content types requires ongoing research and development.

- Implementing effective deep fake detection often requires significant computational resources. This can be a barrier for smaller platforms or organizations with limited resources, limiting their ability to deploy advanced detection systems.

- Addressing these challenges requires a multidisciplinary approach, involving advancements in machine learning, computer vision, and ongoing collaboration between researchers, industry stakeholders, and policymakers.

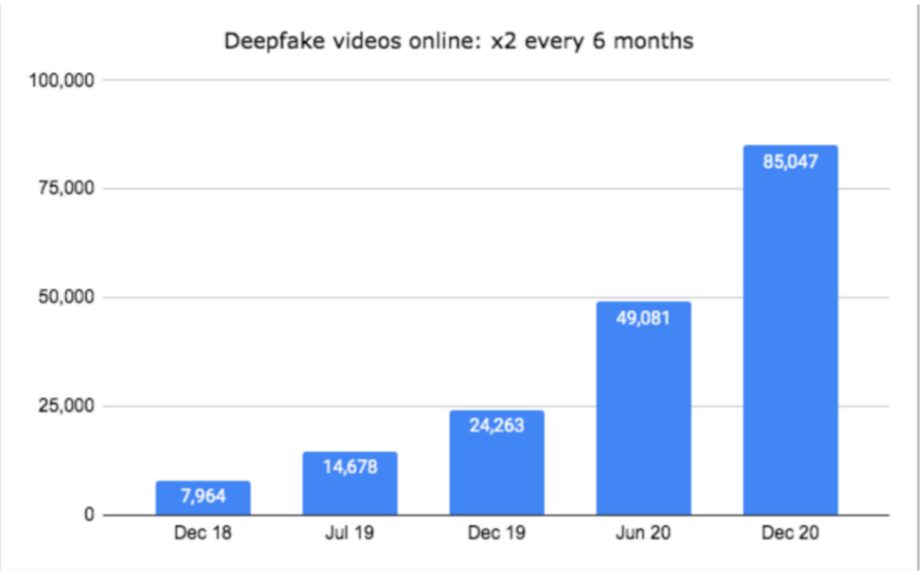


Fig 1.1.1: Rapidly advancing deepfakes

1.2 IDENTIFICATION OF PROBLEM

- In this era of technology of leaving digital footprints, with the rapid process in fields of Artificial Intelligence and deep learning mechanisms, the manipulation has taken a step forward in the case of multimedia. Though this manipulation has its applications in areas of entertainment, education etc. but malicious users have used it for unlawful, illegitimate applications.

- The spreading of misinformation and formatting of videos has created ruckus amongst the people. The people doing this have tried to manipulate police records and evidence, political records and statements and have harassed or blackmailed the victims of the situation.

- Manipulated, doctored high-quality and realistic videos have recently become known as Deepfakes. This false representation of originality has also led to replacement of one's image to another for the purpose of stealing their identity or for defaming them. This can cause severe issues with data security and protection. Creating fake images and videos using image processing tools such as GNU Gimp and Adobe Photoshop is a big problem.

- They are the main source of fake news and are often used to incite crowds, for example. Before acting on a false image, we must check its reality.

- Deepfakes, which are viral in nature, have the potential to negatively stir the minds of millions of people, making their detection a very serious problem. Recent advances in architectures have made generating deep fakes much easier, requiring only a source image and a series of deliberate distortions to generate a believable manipulated image.

- Furthermore, we are currently living in a "post-truth" era where malicious actors use information or disinformation to manipulate public opinion.

- Disinformation is an aggressive tool that can cause serious harm, including election manipulation, creating warmongering situations, and defaming individuals. This problem needs to be taken care of seriously for one's identity protection in this cyber world.

1.3 IDENTIFICATION OF TASKS

Data Collection and Cleaning:

The initial phase involves the meticulous collection of a diverse dataset encompassing both authentic and manipulated media. This dataset serves as the foundation for training and evaluating the detection system. Subsequently, a rigorous data cleaning process is implemented to ensure the integrity and reliability of the dataset, removing any biases or inconsistencies that might impact the effectiveness of the deepfake detection models.

Feature Extraction:

Once the dataset is prepared, the next step is to extract relevant features from the media under analysis. Depending on the type of content, this could involve extracting facial landmarks from images or processing text transcripts for video or audio content. These features serve as input parameters for the subsequent machine learning models, capturing essential information for discerning between genuine and manipulated media.

Machine Learning Model Training:

The extracted features are utilized to train machine learning models, often employing sophisticated techniques like deep neural networks. The training process involves using a labeled dataset, where each piece of media is categorized as either genuine or manipulated. The machine learning models learn to recognize patterns and distinctions that differentiate between authentic and deepfake content, enhancing their ability to make accurate predictions.

Visual Cues Examination:

A crucial aspect of deepfake detection involves a detailed examination of visual cues within the media. This includes scrutinizing artifacts, irregularities, and inconsistencies in texture, lighting, and shadows present in images and videos. These visual cues often reveal subtle anomalies that might be indicative of manipulation. Advanced algorithms are employed to analyze these visual features, contributing to a more comprehensive understanding of the authenticity of the content.

Interpretability and Transparency:

To enhance the reliability of deepfake detection systems, efforts are directed toward developing methods that interpret and explain the decisions made by the underlying deep learning models. This interpretability not only provides transparency into the decision-making process but also fosters trustworthiness. Users and stakeholders benefit from understanding how the system arrives at its conclusions, contributing to increased confidence in the system's efficacy.

Continuous System Update and Adaptation:

Recognizing the dynamic nature of deepfake techniques and emerging threats, a critical step in the deepfake detection process is the continuous update of the detection system. This involves incorporating new data, adapting the models to evolving deepfake methods, and staying ahead of potential challenges. Regular updates ensure that the detection system remains resilient and effective in the face of rapidly advancing deepfake technologies.

By embracing a comprehensive approach that spans data preparation, feature extraction, machine learning model training, visual cues examination, interpretability, and continuous adaptation, deepfake detection systems are better equipped to handle the intricacies of identifying manipulated media in an ever-changing landscape.

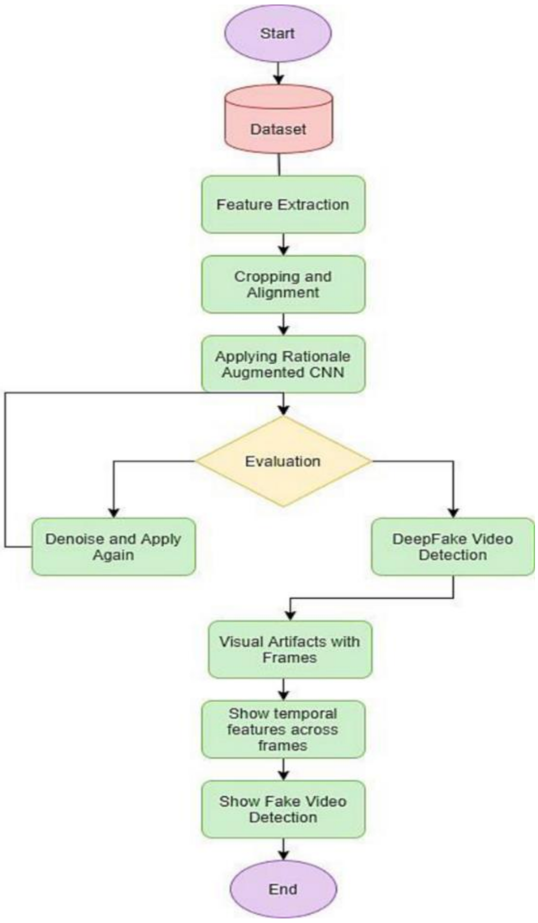
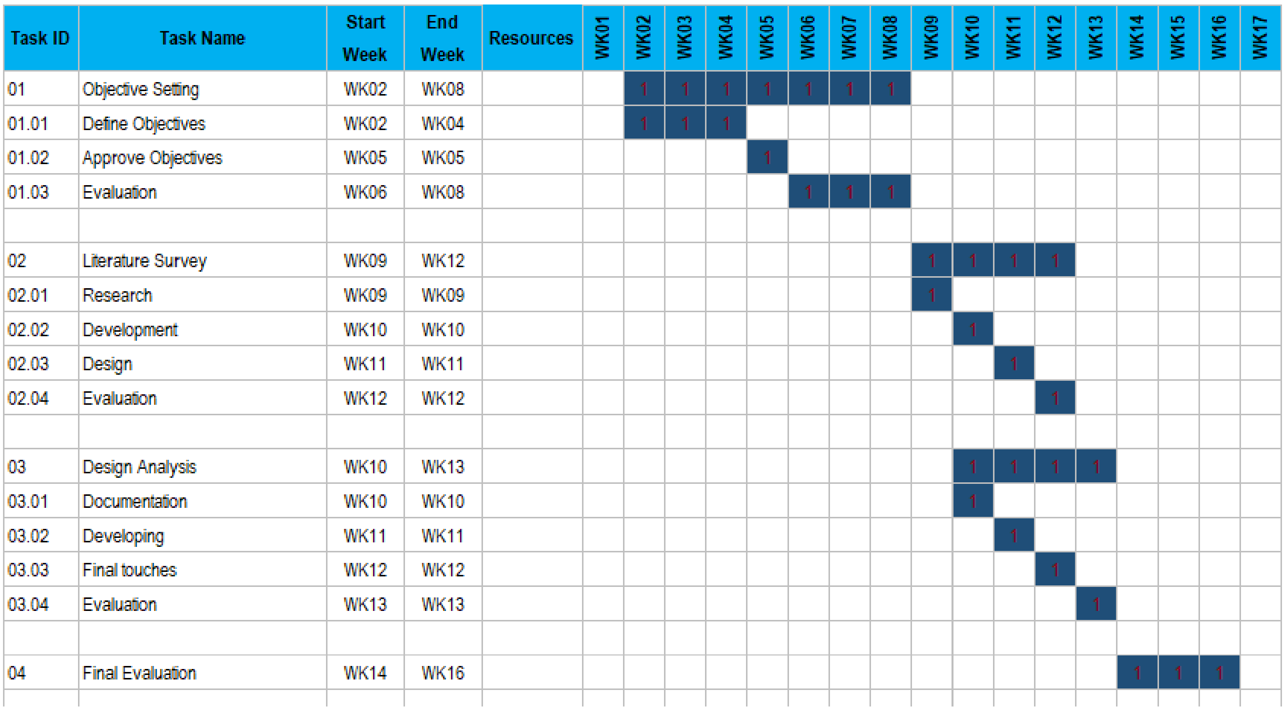

Start
Dataset
Feature Extraction
Cropping and Alignment
Applying Rationale Augmented CNN
Evaluation
Denoise and Apply Again
DeepFake Video Detection
Visual Artifacts with Frames
Show temporal features across frames
Show Fake Video Detection
End

Fig 1.3.1: Flowchart of tasks

1.4 Timeline



Task ID	Task Name	Start Week	End Week	Resources	WK01	WK02	WK03	WK04	WK05	WK06	WK07	WK08	WK09	WK10	WK11	WK12	WK13	WK14	WK15	WK16	WK17
01	Objective Setting	WK02	WK08			1	1	1	1	1	1	1									
01.01	Define Objectives	WK02	WK04			1	1	1													
01.02	Approve Objectives	WK05	WK05						1												
01.03	Evaluation	WK06	WK08							1	1	1									
02	Literature Survey	WK09	WK12										1	1	1	1					
02.01	Research	WK09	WK09										1								
02.02	Development	WK10	WK10											1							
02.03	Design	WK11	WK11												1						
02.04	Evaluation	WK12	WK12													1					
03	Design Analysis	WK10	WK13											1	1	1	1				
03.01	Documentation	WK10	WK10											1							
03.02	Developing	WK11	WK11												1						
03.03	Final touches	WK12	WK12													1					
03.04	Evaluation	WK13	WK13														1				
04	Final Evaluation	WK14	WK16															1	1	1	

Fig 1.4.1: Gantt Chart

ORGANIZATION OF THE REPORT

In this project report, we plan to introduce and explain a model for deep fake detection for effective use of AI and deep learning models:

Introduction:

In this section, we aim to delve into the prevalent methods and challenges associated with deepfake technology, emphasizing the critical need for an effective solution. By identifying the broad problem of deepfake proliferation and its potential for misinformation and fraud, the introduction sets the stage for proposing a robust deep fake detection model. The justification for the study lies in the urgency to safeguard digital content authenticity and credibility. The timeline outlined provides a roadmap, illustrating the projected milestones and phases of the project, ensuring a structured and efficient execution.

Literature Review:

This section involves a comprehensive exploration of existing literature worldwide, highlighting the historical evolution of deepfake detection challenges. It pinpoints when the problem was initially identified and draws upon documentary evidence to substantiate claims. A critical analysis of earlier solutions serves as a foundation for identifying gaps and opportunities. The goals and objectives of the project are meticulously outlined, aligning with the identified shortcomings in the existing approaches. By synthesizing insights from various sources, the literature review provides a nuanced understanding of the landscape, guiding the development of an innovative deep fake detection model.

Design Flow Process:

The design flow process elucidates the features and intricacies of the proposed solution. It not only presents the core components of the deep fake detection model but also explores alternative designs and processes, offering a comprehensive view of the decision-making process. Flowcharts visually depict the sequential steps and interactions within the proposed solution, providing clarity on the work/design flow. This section aims to equip readers with a detailed understanding of the model's architecture, ensuring transparency in its functionality.

Result Analysis and Validation:

Data validation is a pivotal aspect of ensuring the reliability and accuracy of the proposed deep fake detection model. This section dives into the methodologies employed for validating the data and meticulously preparing the project report. Design drawings and solid models further illustrate the tangible outcomes of the project, offering visual representations of the model's efficacy. The rigorous analysis of results includes insights into the model's performance, its ability to accurately detect deepfakes, and any challenges encountered during the validation process.

Conclusion and Future Work:

In the concluding section, the report synthesizes the expected results, outcomes, and any deviations from initial projections. It provides a comprehensive overview of the effectiveness of the proposed deep fake detection model and suggests modifications based on the observed results. The conclusion serves as a reflection on the project's achievements and potential areas for improvement. Future work is outlined, presenting opportunities for extending and refining the proposed solution. This forward-looking perspective encourages ongoing innovation and adaptation to stay ahead of evolving challenges in the realm of deepfake technology.

CHAPTER 2

LITERATURE SURVEY

2.1 TIMELINE OF THE REPORTED PROBLEM

In the era of advanced artificial intelligence and deep learning, the term "deepfake" has emerged as both a technological marvel and a growing concern. In the past two decades, the trend of creating fake media using deepfake has increased rapidly. These deepfakes can be used to put a different image on an image of someone else. This makes use of deep learning models, particularly GAN's. "The 21st centuries' answer to photoshopping, deepfakes use a form of artificial intelligence called deep learning to make images of fake events, hence the name deepfake (deep learning + fake)."

Here are some important incidents which led to the spread of deepfakes, and which made people more familiar to it:

- **Reddit's deepfakes – 2017:** It was one of the first online communities that drew the world's attention towards the newly made fake videos. It was mostly used to make AI-generated videos of celebrities using face swaps.

- **Fake Barack Obama video - 2018**: In April 2018, a fake video of former US president Barack Obama was created using deepfake. The video was a public announcement, and it went viral.

- **Pornography – 2018:** The spread of deepfake technology led to its illegal use by forgers to create adult videos using popular actor and actresses' face. This incident raised privacy concerns regarding the misuse of this technology.

- **Deepfake apps – 2020:** After the widespread misuse of deepfakes to cate videos, it reached another level with the creation of user-friendly deepfake apps. These apps made it user for non-tech people to use deepfake to create forged videos and images.

- **Tom Cruise Deepfake - 2021:** In early 2021, a deepfake video featuring actor Tom Cruise went viral on social media. This video was so convincing that it sparked discussions about the potential for deepfakes to be used in the entertainment industry.

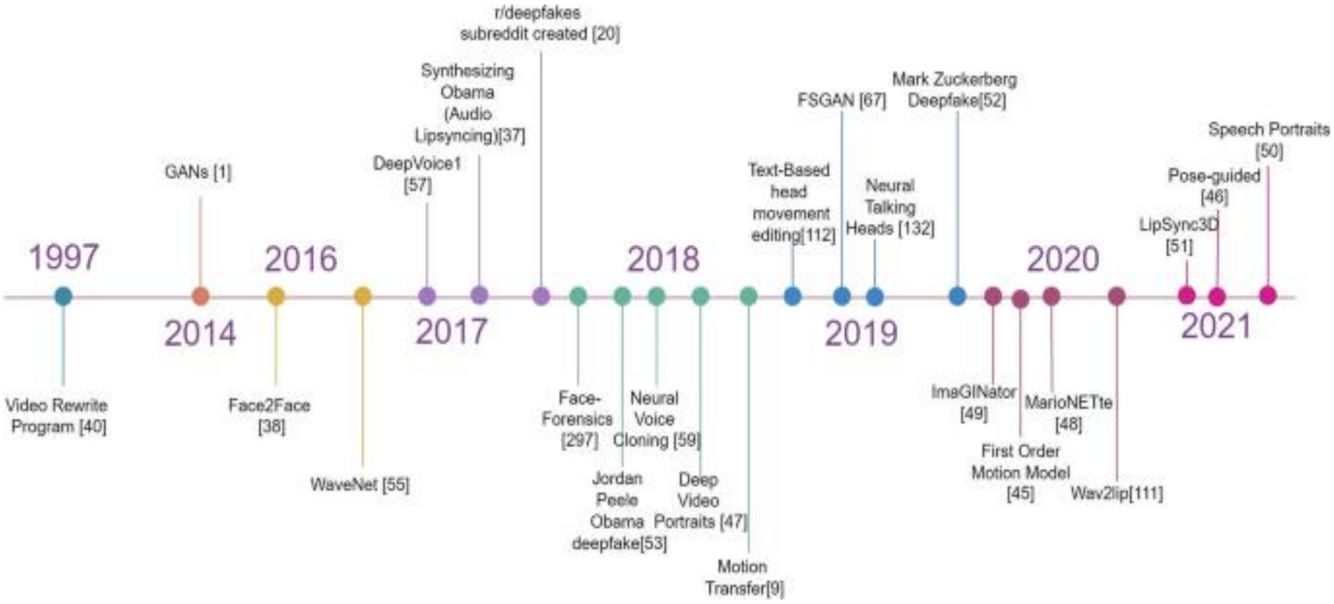

1997
Video Rewrite Program [40]
GANs [1]
2014
2016
Face2Face [38]
WaveNet [55]
DeepVoice1 [57]
2017
Synthesizing Obama (Audio Lipsyncing)[37]
r/deepfakes subreddit created [20]
Face-Forensics [297]
Jordan Peele Obama deepfake[53]
Neural Voice Cloning [59]
2018
Deep Video Portraits [47]
Motion Transfer[9]
Text-Based head movement editing[112]
FSGAN [67]
2019
Neural Talking Heads [132]
Mark Zuckerberg Deepfake[52]
ImaGINator [49]
First Order Motion Model [45]
MarioNETte [48]
2020
Wav2lip[111]
LipSync3D [51]
Pose-guided [46]
2021
Speech Portraits [50]

Fig 2.1.1: Timeline of deepfake.

2.2 PROPOSED SOLUTIONS

After several incidents that made the world familiar with deepfakes, it became even more dangerous. Researchers and scholars began to find solutions to combat the misuse of this technology. In this section, we are going to discuss some of the popular proposed solutions.

The first move was to develop and deploy algorithmic tools to differentiate between fake and original media. These were made using machine learning and AI to analyze the media’s signs of manipulation. Various cryptographic methods like encryption of the sensitive media and digital signatures were used to verify the authenticity of media. Moreover, invisible or semi-visible watermarks were embedded into the media to prove their authenticity. Researchers made use of various deep learning and machine learning models like CNN to identify the fake media. Other technologies used were general adversarial networks, RNN, auto-encoder, Celeb-DF3.

Though there is no accurate model to identify deepfake media, scholars are still working on it. Several other methods like that of CNN and GAN have proved to be more accurate than others and they are widely used for the purpose. Making people aware of this technology is currently more important to prevent its spread so that they can identify the manipulated content by themselves and do not rely on heavy technologies.

Keeping in mind all these points, social media and content-sharing platforms can implement policies that prohibit the distribution of malicious deepfake content and provide mechanisms for users to report such content.

2.2.1 Face-Swap Generation

In face-swap, or face replacement, the face of the person in the source video is automatically replaced by the face in the target video, as shown in Fig. 2.2.1.1. Traditional face-swap approaches generally involve a sequence of three steps for performing a face-swap operation.

First step: These tools employ face detection algorithms to identify the face in source images and then select a candidate face image from a facial library that closely resembles the input facial appearance and poses.

Second step: The method replaces specific facial features such as eyes, nose, and mouth from the candidate face onto the source face. This is followed by adjustments to the lighting and color of the candidate face image to match the appearance of input images, ensuring a seamless blend between the two faces.

Third step: The blended candidate replacement is ranked by computing a match distance over the overlap region, providing a quantitative measure of the integration quality. This ranking aids in selecting the most convincing face swap by considering factors like alignment, color consistency, and overall visual coherence.

It's worth noting that traditional face-swap methods have been enhanced with the advent of deep learning techniques, particularly generative models like GANs (Generative Adversarial Networks), which offer more advanced approaches to face swapping. These newer methods demonstrate improved performance and can handle a broader range of facial variations and expressions, contributing to the ongoing evolution of face-swap technology.

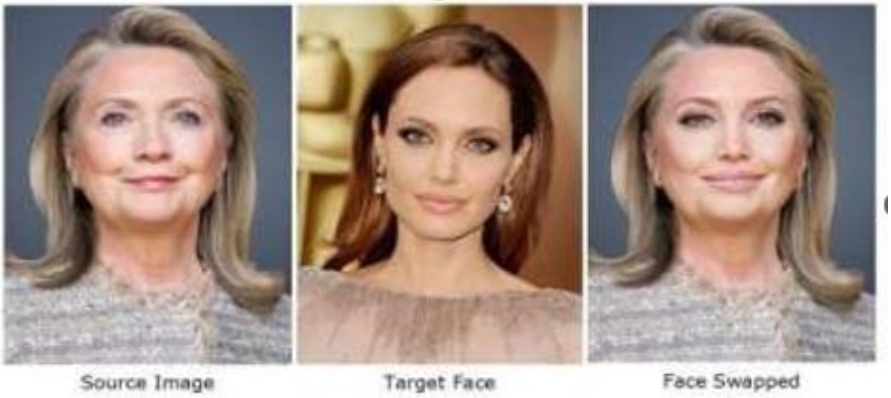


Fig 2.2.1.1: Visual representation of face swap

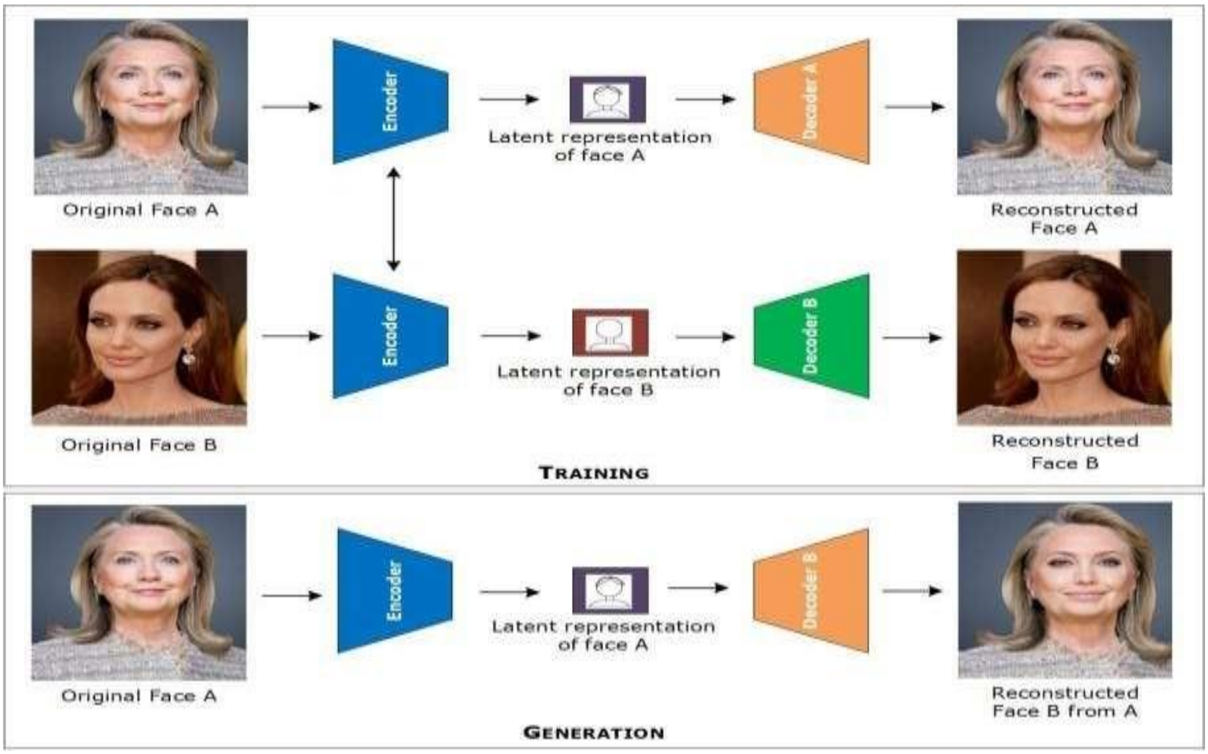


Fig 2.2.1.2: Creation of a Deepfake using an auto-encoder and decoder.

2.2.2 Lip-syncing Generation

Lip-Syncing Approach for Target ID Video Synthesis:

The lip-syncing approach aims to synthesize the target ID video, ensuring that the mouth region in the manipulated video aligns seamlessly with any provided audio input (refer to Fig. 2.2.2.1). The effective synthesis of visual language hinges on accurately capturing the movement and appearance of the lower mouth and surrounding areas, allowing for the creation of lifelike lip movements that complement facial expressions.

First Step: Frame Reselection and Emotional Context Integration:

Initiating the lip-sync process involves the careful reselection of frames from a video or transcription. This selection process takes into account not only the linguistic content but also the emotional nuances of the target message. By incorporating information about the desired emotions, the synthesis strives to produce lip movements that are not only accurate but also emotionally resonant. This step forms the foundation for conveying messages effectively and naturally.

Second Step: RNN-Based Mapping of Audio Features to Mouth Shape:

A pivotal aspect of the lip-syncing approach is the utilization of a recurrent neural network (RNN)-based model. This model is designed to learn the intricate mapping between audio features and the corresponding mouth shape for every frame in the selected sequence. By understanding the correlation between the spoken words and the visual representation of mouth movements, the RNN enhances the accuracy and realism of the lip-syncing process. Frame reselection is then employed to populate the texture around the mouth, utilizing landmarks for precise alignment.

Third Step: Comprehensive Synthesis and Realism Enhancement:

In the final step of the lip-syncing process, synthesis is performed on the lower facial regions, encompassing the mouth, chin, nose, and cheeks. Special attention is given to smoothing the jaw location to ensure a natural flow of movement. Additionally, the video is re-timed to align vocal pauses or other talking head motions, further enhancing the natural appearance of the synthesized content. The holistic approach to synthesis, involving multiple facial regions and meticulous timing adjustments, contributes to the creation of videos that not only appear realistic but also authentically convey the intended message.

The continuous advancement of machine learning techniques, especially in the realm of neural networks, plays a crucial role in refining and optimizing lip-syncing methodologies, allowing for increasingly accurate and expressive visual language synthesis.

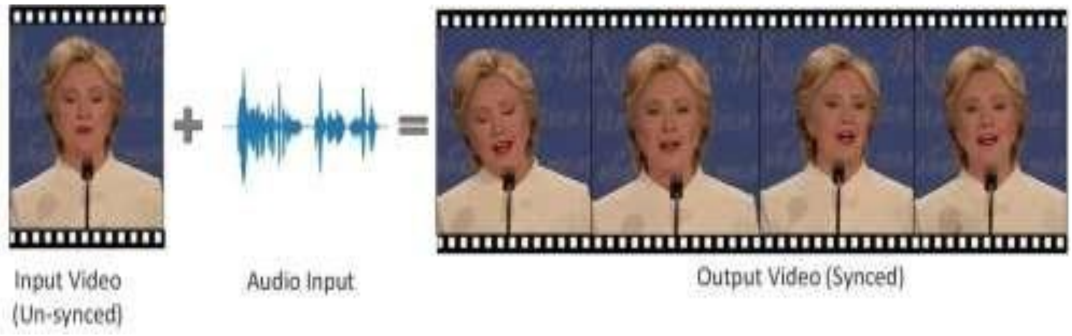


Figure 2.2.2.1 A visual representation of lip-syncing

2.2.3 Puppet-master Generation

Puppeteer: Facial Reproduction for Expressive Manipulation:

Puppeteer, a variant of deepfake technology known as facial reproduction, specializes in manipulating a person's facial expressions by transferring nuanced gestures, eye movements, and head motions to an output video, mirroring those of the source actor. The Puppet Mastery technique aims to generate synthetic content by skillfully transforming human mouth movements, resulting in realistic and expressive facial manipulations.

First Step: 3D Model Tracking and Reconstruction:

The process begins with the use of a standard RGB-D sensor for meticulous tracking and reconstruction of the three-dimensional models of both the source and target actors. This involves capturing detailed facial features, creating a digital representation that forms the basis for subsequent manipulations. The RGB-D sensor captures color information and depth, enhancing the accuracy of the 3D model reconstruction.

Second Step: Deformation Application and Blending:

For each frame in the video sequence, the tracked deformations of the source actor's face are applied to the 3D model of the target actor. This ensures accurate transfer of expressive facial movements, including mouth gestures, from the source to the target. Subsequently, the altered face is seamlessly blended onto the original target face while preserving the facial appearance of the underlying 3D model. The blending process is critical for maintaining visual consistency and ensuring that the manipulated facial expressions harmoniously integrate with the target actor's overall appearance.

Advanced Techniques:

Puppeteer often incorporates advanced techniques to enhance the realism of manipulated facial expressions. This may involve detailed texture mapping, dynamic lighting adjustments, and realistic shading to achieve a convincing and lifelike result. The system may also adapt to variations in lighting conditions and facial poses, contributing to the adaptability and robustness of the Puppeteer technique across diverse scenarios.

Ethical Considerations:

Similar to any deepfake technology, Puppeteer raises ethical considerations regarding its potential misuse for deceptive purposes. The creation of highly realistic facial manipulations necessitates responsible deployment and careful consideration of the ethical implications associated with the technology.

Ongoing research in computer vision and machine learning plays a role in refining Puppeteer and similar techniques, ensuring a balance between technological innovation and ethical considerations in the realm of facial expression manipulation.

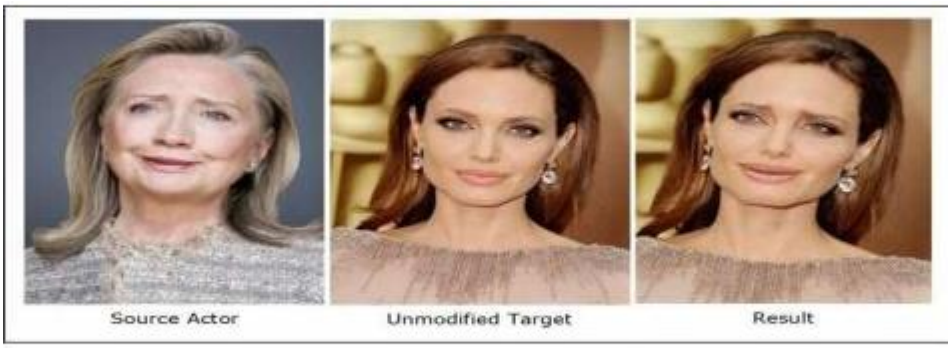


Figure 2.2.3.1: A visual representation of face-reenactment

2.3 BIBLIOMETRIC ANALYSIS

The surge in deepfake publications underscores the urgency in addressing the potential threats posed by manipulated audio and video content. Researchers have made commendable strides in developing detection systems that are both robust and accurate. These advancements are crucial in fortifying the digital landscape against the proliferation of deceptive deepfake content. As the technology behind deepfakes becomes more sophisticated, the demand for detection systems capable of navigating this complexity has intensified.

Challenges and Ongoing Research:

While progress has been notable, the challenges in deepfake detection persist. The dynamic nature of deepfake technology necessitates the continual development of more robust systems. Researchers are actively exploring avenues to enhance the generalization capabilities of detection systems, enabling them to adapt to new and previously unseen data. The computational efficiency of these systems is also a focal point, as real-time deployment becomes increasingly vital in the battle against rapidly disseminating deepfake content.

Multimodal Detection Approaches:

The evolution of deepfake creation techniques across various media formats—videos, images, and audio—has spurred the need for comprehensive, multimodal detection systems. Researchers are diligently working towards creating unified frameworks capable of detecting deepfakes irrespective of the media type. This holistic approach aims to address the multi-faceted nature of the deepfake challenge and ensure a more comprehensive defense against deceptive content.

Global Collaboration and Emerging Trends:

The global distribution of researchers contributing to deepfake detection research is indicative of the collaborative effort required to tackle this pervasive issue. The emerging trends suggest a collective awareness of the evolving landscape, emphasizing the importance of international cooperation in devising effective countermeasures. As the field matures, the need for interdisciplinary collaboration becomes even more pronounced, with experts from diverse backgrounds pooling their expertise to confront the multifaceted challenges posed by deepfake technology.

Real-time Detection and Continuous Adaptation:

The dynamic nature of deepfake evolution demands real-time detection methods. Researchers are actively exploring adaptive techniques that can keep pace with the rapid advancements in deepfake creation. The ability to detect and mitigate deepfake threats in real time is paramount in minimizing their potential impact on public perception, trust, and overall digital security.

In conclusion, the bibliometric analysis not only highlights the remarkable growth in deepfake detection research but also underscores the necessity for continuous innovation and collaboration to stay ahead of the ever-evolving landscape of deepfake technology. Researchers worldwide are at the forefront of this effort, contributing to the development of cutting-edge solutions that safeguard the authenticity and credibility of digital content.

2.4 REVIEW SUMMARY

Commonly used as a source for spreading false information to the public via social sites to mislead, cause emotional distress, or deliberately influence decisions, policies, and actions image. As a result, the authenticity of digital images has recently become a major research area in the literature. The first explainable fake news detector, the XFake system, is presented to help end users detect whether a message is fake or not.

Context of Image Forgery in the Digital Age:

In the era of digital communication, the manipulation of images has emerged as a potent tool for disseminating false information across social platforms. These manipulations are often orchestrated to mislead, cause emotional distress, or exert influence on decisions, policies, and actions. Consequently, the study of digital image authenticity has gained prominence in contemporary literature. Researchers are actively exploring innovative methods to address the challenges posed by the proliferation of manipulated images, particularly in the context of spreading misinformation.

Introduction of XFake System:

In response to the pressing need for effective fake news detection, the XFake system is introduced as the first explainable fake news detector. This system is designed to empower end-users in discerning the authenticity of messages circulating on social platforms. By providing transparency in its decision-making process, the XFake system represents a pivotal step in countering the spread of false information and enhancing digital literacy among users.

Active vs. Passive Approaches in Image Forgery Detection:

Image forgery detection strategies are broadly classified into active and passive approaches. Active methods leverage digital watermarks or signatures embedded in the original image to verify its authenticity.

However, this approach is constrained by the necessity for authorized personnel or the capture device itself to carry out the embedding process. On the other hand, passive approaches have garnered attention due to their independence from additional information, making them more versatile and applicable in various scenarios.

Passive Approaches: Copy-Motion and Image Splicing Techniques:

Within passive approaches, the literature highlights copy-motion and image splicing techniques as promising avenues for image forgery detection. Copy-motion involves duplicating and relocating portions of an image, concealing vital features, duplicating objects, or altering the image's meaning. Image splicing, akin to fake transitions, stitches together multiple images to create a deceptive composite. These techniques play a crucial role in the arsenal of passive forgery detection methods, offering insights into identifying manipulated content without relying on embedded information.

Challenges in Deepfake Detection:

The realm of deepfake detection introduces additional complexities. Existing methods often target specific types of deepfakes or employ singular strategies, limiting their effectiveness in the face of diverse deepfake types and generation methods. Evading detection becomes a concern, prompting the need for a more robust and versatile detection strategy.

Proposed Comprehensive Deepfake Detection Method:

To address the challenges posed by the wide variety of deepfake types and evasion tactics, the proposed method adopts a multifaceted approach. By combining various detection strategies into a unified model, this method aims to enhance the power and accuracy of deepfake detection. The integration of diverse strategies not only broadens the scope of detection but also increases the resilience of the model against evolving deepfake techniques.

In summary, the evolving landscape of image forgery, especially in the context of spreading misinformation, calls for innovative solutions like the XFake system and a comprehensive approach to deepfake detection. Researchers continue to explore new avenues to fortify the authenticity of digital content and mitigate the potential consequences of manipulated images in the digital age.

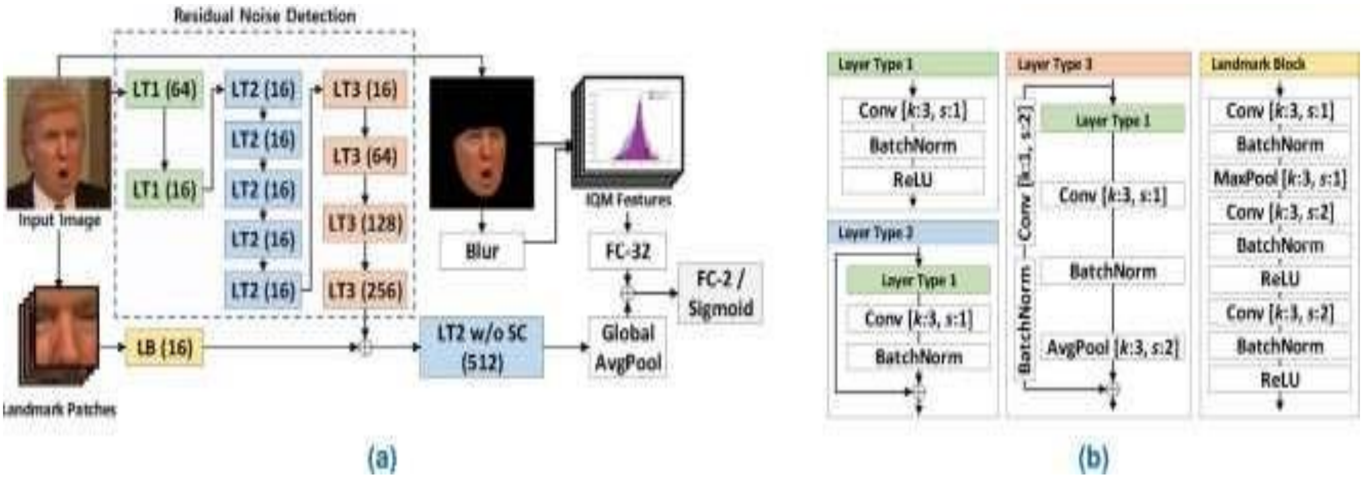


Figure 2.4.1 Proposed network architecture for deepfake detection.

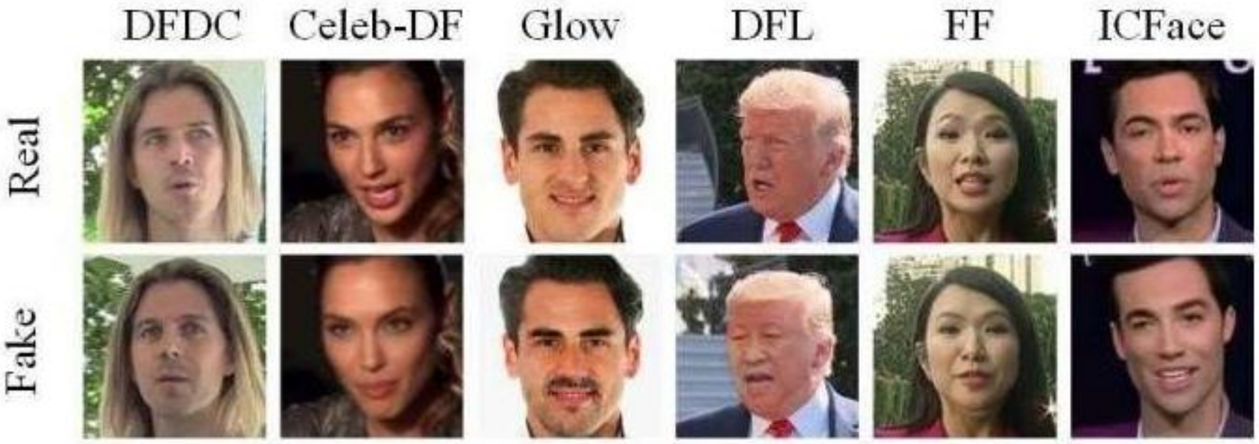


Figure 2.4.2 Some Dataset Images

2.5 PROBLEM DEFINITION

In the dynamic landscape of digital content, the urgency to deploy effective deepfake detection systems has become paramount. These systems serve as a critical defense against the proliferation of deceptive, AI-generated multimedia content that convincingly replicates genuine human actions and speech in videos and audio recordings. The central challenge is to construct a system that not only accurately but also efficiently identifies and flags deepfake videos and images, discerning them from authentic media.

The primary objective of a robust deepfake detection system is to navigate the intricate landscape of manipulated content and distinguish it from its authentic counterpart. The inherent difficulty lies in the ever-increasing sophistication of deepfake techniques, making them progressively elusive and challenging to identify. Researchers are actively engaged in the development of innovative solutions that go beyond the limitations posed by evolving deepfake technology.

The multifaceted nature of the problem demands a comprehensive approach, as deepfakes continue to evolve in complexity. A successful deepfake detection system must meticulously scrutinize various features and inconsistencies within the media it analyzes. This includes a detailed examination of facial expressions, subtle lighting variations, and anomalies in the audio component. By integrating these elements into the detection process, the system enhances its capacity to discern between authentic content and meticulously crafted deepfakes.

The challenges in deepfake detection underscore the importance of continuous research and innovation in the field. As deepfake technology evolves, detection systems must adapt to the changing landscape to remain effective. Significantly, researchers are making substantial progress in advancing the accuracy and robustness of deepfake detection systems, contributing to the ongoing arms race between creators of deceptive content and those developing countermeasures.

Effectively addressing these challenges is crucial for mitigating the growing threat posed by deepfake technology. As these manipulations become more prevalent and sophisticated, the development and deployment of resilient deepfake detection systems play a pivotal role in upholding the integrity of digital content, safeguarding against misinformation, and preserving trust in multimedia communication channels. The continuous evolution of deepfake detection strategies will be instrumental in staying ahead of emerging threats and ensuring the reliability of digital media in the years to come.

2.6 GOALS AND OBJECTIVES

The goals of a deepfake detection system are to:

1. Identify the presence of deepfake content in multimedia.
2. Ensure trust in digital media.
3. Protect people from the harmful effects of deepfakes, such as misinformation, disinformation, and fraud.
4. Inform users about the existence and risks of deepfakes.

The objectives of a deepfake detection system are to:

1. Develop a system that can accurately and efficiently identify and flag deepfake videos and images.
2. Develop techniques for detecting deepfakes.
3. Develop a system that can generalize to new and unseen data.
4. Continuously update and enhance the system to address new techniques and challenges.

CHAPTER 3

DESIGN FLOW / PROCESS

3.1. Evaluation & Selection of Specifications/Features

3.1.1. Objectives

The objective of this Project is to explore the emerging nature of Internet content while doing so categorize and detecting videos and images as real or Deepfake. In doing so we must give the confidence level of our prediction based on the model trained using the publicly available Dataset.

Understand the Problem:

Consider the broader implications of Deepfakes, including their potential impact on misinformation and privacy. Stay updated on the latest advancements in Deepfake technology to inform your model development.

Collect and Preprocess Data:

Ensure that the dataset includes variations in lighting conditions, camera angles, and facial expressions to enhance the model's robustness. Implement techniques to handle potential biases in the data, as biased datasets can lead to biased models.

Data Splitting:

Stratify your dataset during the splitting process to maintain the distribution of real and Deepfake samples in each subset.

Select a Model Architecture:

Experiment with pre-trained models such as ResNet, VGG, or EfficientNet, and fine-tune them for your specific task. Consider incorporating temporal information for video analysis, possibly using 3D CNNs or recurrent neural networks (RNNs).

Model Training:

Explore transfer learning to leverage knowledge gained from a model trained on a large dataset for a similar task.Regularly monitor training curves to detect signs of overfitting and adjust the learning rate accordingly.

Model Evaluation:

Use confusion matrices and ROC curves to gain a more comprehensive understanding of your model's performance. Investigate misclassifications to identify patterns and potential areas for improvement.

Test the Model:

Conduct a thorough analysis of your model's performance on the testing set, considering different evaluation metrics and scenarios.

Confidence Level Calculation:

Implement techniques such as Monte Carlo Dropout or uncertainty estimation to quantify the model's confidence in its predictions.Consider providing users with an interpretability score alongside predictions to enhance transparency.

Model Interpretability:

Utilize tools like LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) to provide insights into individual predictions. Ensure that your interpretation methods align with the expectations and requirements of end-users and stakeholders.

Deployment:

Choose a deployment strategy that aligns with the project's scale and requirements, whether it's edge deployment for real-time processing or cloud-based solutions for batch processing. Implement security measures to protect the model from adversarial attacks.

Continuous Improvement:

Establish a pipeline for regularly updating your model with new data and retraining it to adapt to evolving Deepfake techniques.Engage with the research community to stay informed about the latest developments in Deepfake detection and mitigation.

Ethical Considerations:

Consider implementing safeguards to prevent potential misuse of your model. Collaborate with ethicists, legal experts, and other stakeholders to address ethical concerns and ensure responsible AI development.By incorporating these considerations, your project will be better equipped to tackle the challenges associated with detecting Deepfake content on the internet.

3.1.2. Lifespan

Since This Machine Learning project is divided into three Phases, namely Training, making it available using a website, and testing of the model on unrecognized data, the time required to train the model as well as the website is 60 days, and the Testing is estimated to be completed in 15 days.

1. Comprehensive Training Phase:

The delineation of the machine learning project into three distinct phases—Training, Deployment via a Website, and Testing on Unrecognized Data—entails a meticulous and well-structured approach. The initial Training Phase is a cornerstone, requiring an estimated duration of 60 days. This duration encapsulates not only the model's training on diverse datasets but also incorporates the crucial steps of hyperparameter tuning, optimization, and iterative refinement.

2. Deployment via Website:

The transition from the Training Phase to Deployment marks a pivotal juncture. The model's availability via a dedicated website is a testament to the project's practical applicability and accessibility. The duration allocated for this deployment phase underscores the importance of a smooth transition from model training to real-world application, and it is anticipated to span the initial 60 days.

3. Rigorous Testing on Unrecognized Data:

The Testing Phase represents the culmination of the project's lifecycle, where the trained model is rigorously tested on previously unseen, unrecognized data. This phase is earmarked for an estimated duration of 15 days, emphasizing the need for thorough evaluation, validation, and performance assessment. The brevity of this testing period is intentional, reflecting the project's efficiency and the model's readiness to navigate real-world scenarios promptly.

Holistic Lifespan Considerations:

Beyond the individual durations assigned to each phase, it's essential to view the project's lifespan holistically. Continuous monitoring, updates, and adaptation are integral aspects of the project's ongoing journey. Lifespan considerations encompass not just the formalized phases but also encompass the potential for future enhancements, advancements in deepfake detection techniques, and the evolving landscape of multimedia manipulation.

Adapting to Emerging Challenges:

As technology evolves and new challenges emerge, the project's lifespan extends beyond the initially outlined phases. Provisions for adapting to emerging challenges, incorporating feedback from real-world usage, and staying current with the latest developments in machine learning and deepfake detection technologies are inherent in the project's adaptable design. This adaptability ensures that the deepfake detection system remains effective, even in the face of evolving threats and manipulation techniques.

Knowledge Transfer and Documentation:

As the project progresses through its lifespan, knowledge transfer and documentation play crucial roles. Ensuring that insights gained during training, deployment, and testing phases are well-documented facilitates seamless knowledge transfer to future iterations or related projects.

In essence, the project's lifespan spans the dedicated durations allocated to each phase, but its impact extends far beyond, embracing adaptability, continuous improvement, and a commitment to addressing emerging challenges in the realm of deepfake detection.

3.1.3. Life Cycle

- The project lifecycle started with devising the idea for the project by starting to search for a problem statement and ideas to solve it.
- After successfully deciding on the project, our team started to research existing Deepfake Detection Technologies and their Features.
- Then we decided on algorithms to use to detect Deepfake.
- After completing the design of the project, the project was built in around 60 days, being tested for any faults, and then it will be deployed as the final step of the development lifecycle.

3.1.4. Risk and Uncertainty

Project development poses a lot of risks and uncertainties which need to be taken care of during the development process. Some of them include:

Mitigation Strategies:

Conduct thorough testing during the development phase to identify and address technical faults early on. Implement continuous integration and automated testing to catch issues as soon as they arise. Regularly update and patch software dependencies to mitigate security vulnerabilities.

Contingency Plans:

Develop contingency plans for potential technical failures, including data corruption, model training issues, or infrastructure failures. Establish a rollback strategy in case of deployment issues and ensure that backups are regularly created and tested.

Documentation:

Maintain comprehensive documentation for the entire development process, including data preprocessing, model architecture, and training procedures. This facilitates easier troubleshooting in case of technical faults. Low Confidence Levels in Some Datasets (Possibly Due to Overfitting):

Data Quality Assurance:

Implement rigorous data quality assurance processes to ensure that the dataset is representative, diverse, and free from biases. Regularly audit the dataset for potential issues and anomalies, and update it as needed.

Regular Model Evaluation:

Continuously monitor your model's performance, especially its ability to generalize to new, unseen data. Use validation sets and metrics to identify and mitigate overfitting issues during the training phase.

Ensemble Learning:

Implement ensemble learning techniques by combining predictions from multiple models. This can help improve generalization and reduce the impact of overfitting from individual models.

Regular Model Retraining:

Schedule regular model retraining intervals to incorporate new data and adapt to evolving trends in Deepfake generation.

Hyperparameter Tuning:

Experiment with different hyperparameter settings during model training to find the optimal configuration that minimizes overfitting.

Cross-Validation:

Use cross-validation techniques to assess the model's performance on multiple subsets of the data. This provides a more robust evaluation and helps identify potential overfitting.

Regular Model Interpretability Analysis:

Analyze the interpretability of your model to understand its decision-making process. This can provide insights into potential overfitting or biases in the training data. By addressing these risks through a combination of proactive measures, contingency planning, and ongoing monitoring, you can enhance the robustness and reliability of your Deepfake detection project. Regular updates to your risk management plan as the project evolves will also contribute to successful risk mitigation.

3.1.5. Directions

- We are working on our project with the direction given by our co-supervisor and supervisor such as time limits, useful resources, and technical help to solve the minor bugs and issues in our project.
- After developing the prototype of the website, we took customer feedback about what the website lacked and where it needed some modifications.
- Based on the feedback received, we worked in the directions provided by them to build a customer-friendly application.

3.2. Design Constraints

3.2.1. Cost

- The project is purely software-based. The development has zero cost involved because we are testing the site in localhost which doesn't require any money, and we are using free hosting to run in development mode.
- Even if the project requires enhancement we might go for funding however it would not be required unless we have a large user base.

3.2.2. Scope

Since The rise of Web Fake content and Deepfakes are playing an increasing role in the political and news domain. There is a lack of objective truth Through this website we plan to clearly distinguish between Digitally edited and real Images.

3.2.3. Quality

To achieve high-quality code, we are using various modern libraries in our codebase Our team has tried our best to keep the quality of the project in check. Rest, the client feedback will allow us to make modifications as required.

3.2.4. Benefits

This project will help users to upload and see whether a specific video is doctored or real. It will also help to educate users about the various Deepfakes technologies.

3.3. Analysis and Feature finalization subject to constraints

The main features of the project are:

- Every User can use and detect whether a video is manipulated or real.
- It Counters trolls and Bots that are used to spread misleading content among people.
- Help to tackle conspiracy theories and disinformation spread by politicians and foreign Governments.
- Help Law enforcement take countermeasures against people spreading fake and manipulated media.
- Its user-friendly design, ensuring that every user, regardless of technical expertise, can easily utilize the tool to discern whether a video is manipulated or authentic.
- By detecting and categorizing manipulated videos, the tool aids in identifying and mitigating the impact of orchestrated campaigns that seek to sow discord, spread misinformation, and influence public opinion through deceptive multimedia content.
- Conspiracy theories often thrive on the dissemination of manipulated media. The project takes a proactive stance in tackling conspiracy theories by providing a means to identify and verify the authenticity of videos.
- By offering a tool for detecting fake and manipulated media, the project becomes a valuable resource for countering attempts to manipulate public opinion, protect democratic processes, and uphold the integrity of information shared by political entities.
- A tool that aids in the identification and categorization of manipulated content, law enforcement can take targeted countermeasures against individuals or entities involved in the intentional dissemination of deceptive multimedia content.
- By enabling users to discern between real and manipulated media, it contributes to media literacy and serves as a preventive measure against the potential harmful effects of misinformation.

In summary, the project's main features encompass user-friendly accessibility, mitigation of trolls and bots, combating conspiracy theories, addressing disinformation from politicians and foreign governments, support for law enforcement, and an educational and preventive role. Together, these features position the project as a versatile and impactful tool in the ongoing efforts to uphold the integrity of digital media and protect users from the negative consequences of manipulated content.

3.4 Design Flow

We cast the forgery detection as a binary classification problem of the manipulated videos. This section shows the results of various forgery **Learned features** detection methods.

1. Steganalysis features were cast to a CNN-based network by Cozzolino et al. It was fine-tuned on a large dataset
2. The convolutional neural network uses a constrained convolutional layer followed by two max-pooling, two convolutional, and three fully-connected layers proposed by Bayar and Stamm. To suppress high-level content of image-constrained convolutional layer was specifically designed.
3. Stats-2L network that had the overall best performance among different CNN architectures with global pooling that computes four statistics i.e mean, variance, maximum and minimum. It was proposed by Rahmouni et al
4. A network that has two inception modules and two classic convolution layers interlaced with max-pooling layers. We take the mean squared error between predicted and true values instead of cross-entropy loss. This was inspired by InceptionNet and is known as MesoInception.
5. The task of replacing the final fully connected layer with two outputs is done by a traditional CNN approach based on separable convolutions with residual connections. This was trained on ImageNet and is known as XceptionNet. In this method, all other layers are initially given ImageNet weights. We fix weights up to the last layers and pre-train the network for 3 epochs to set up a fully connected layer. Afterward, we train this network for 15 epochs and choose the model with the highest validation accuracy.
6. An ensemble learning approach was employed to further enhance the robustness and accuracy of forgery detection. This approach involves combining the outputs of multiple detection models to make a final prediction. The ensemble model leverages the diverse strengths of various architectures, improving the overall performance by mitigating individual model biases and increasing the generalization capacity.
7. To capitalize on the wealth of knowledge embedded in pre-trained models, transfer learning was incorporated into the forgery detection framework. Specifically, models like VGG16, ResNet, and MobileNet were fine-tuned on the forgery detection task.
8. Recognizing the importance of real-time forgery detection, a specialized model was developed

with a focus on minimizing inference time while maintaining high accuracy. This model employs optimized architectures and lightweight components to facilitate swift decision-making, making it suitable for applications where timely forgery detection is paramount.

9. In response to the dynamic nature of forgery techniques, a continuous model updating and adaptation strategy were implemented. This involves periodically retraining the models on new data, ensuring that the detection system stays resilient against emerging forgery methods. By staying abreast of evolving trends in digital manipulation, the models remain effective in countering the latest challenges in the realm of forgery detection.

In summary, the forgery detection project employs a diverse array of methods, including ensemble learning, transfer learning, attention mechanisms, real-time detection, continuous model updating, and integration with blockchain technology.

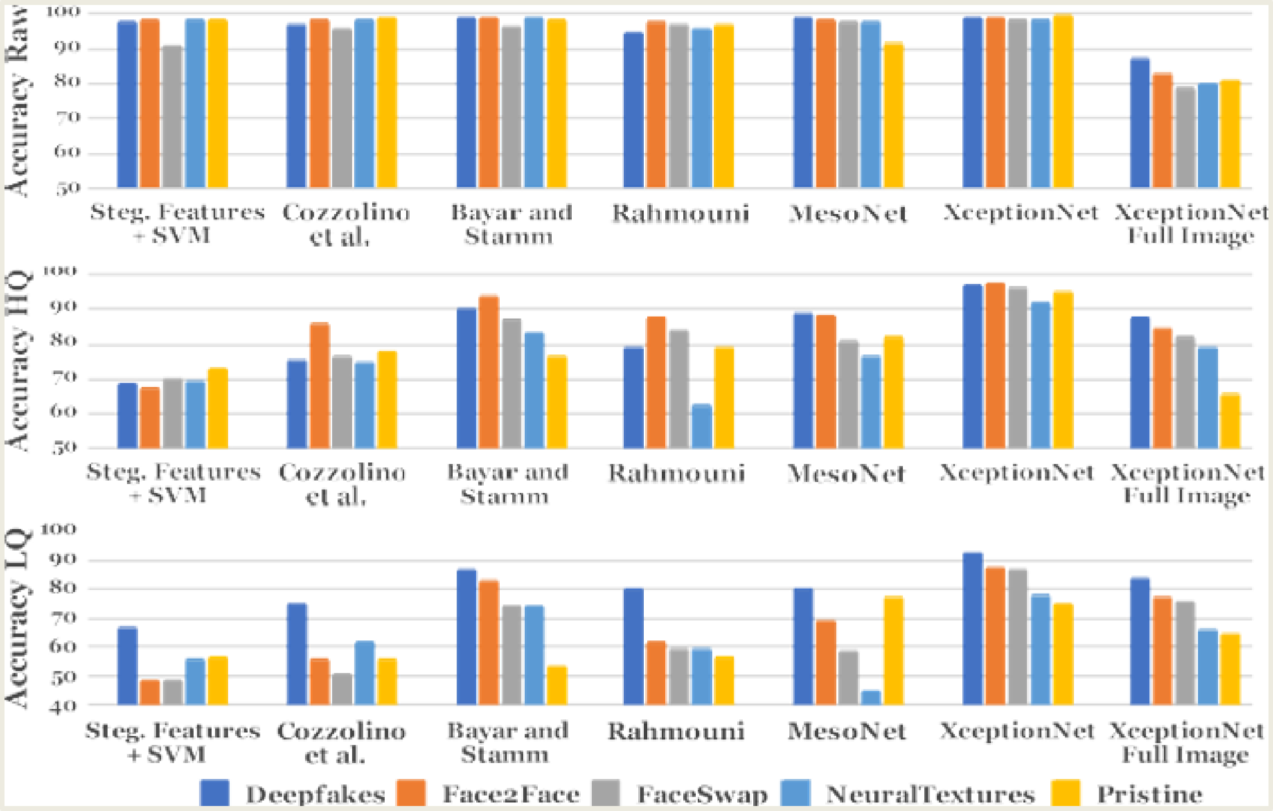


Figure 3.4.1: Binary precision values of our baselines when trained on all four manipulation methods simultaneously.

3.5 Design selection

3.5.1. Use Case Diagram

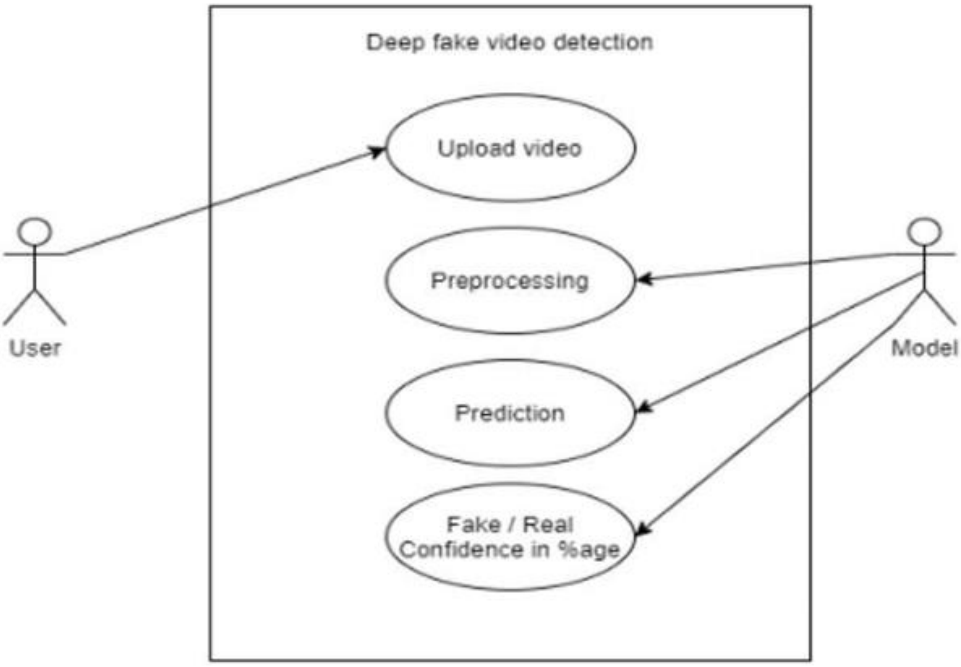


Figure: 3.5.1.1: Use Case Diagram

3.5.2 Data flow diagrams

3.5.2.1 DFD level 0

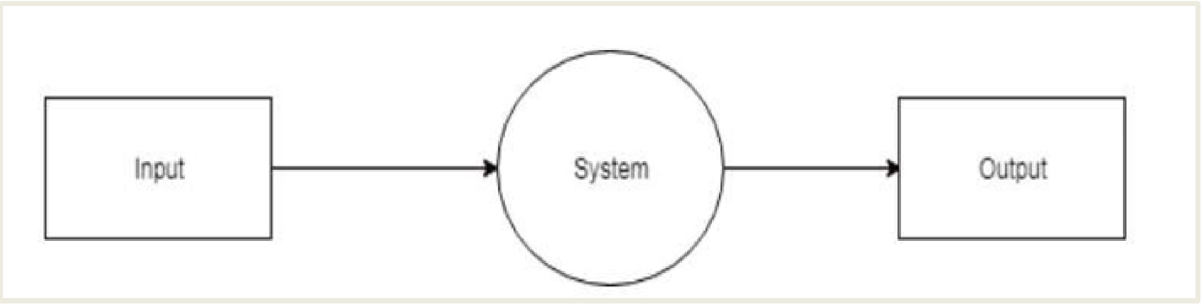


Figure 3.5.2.1: DFD level 0

3.5.2.2 DFD level 1

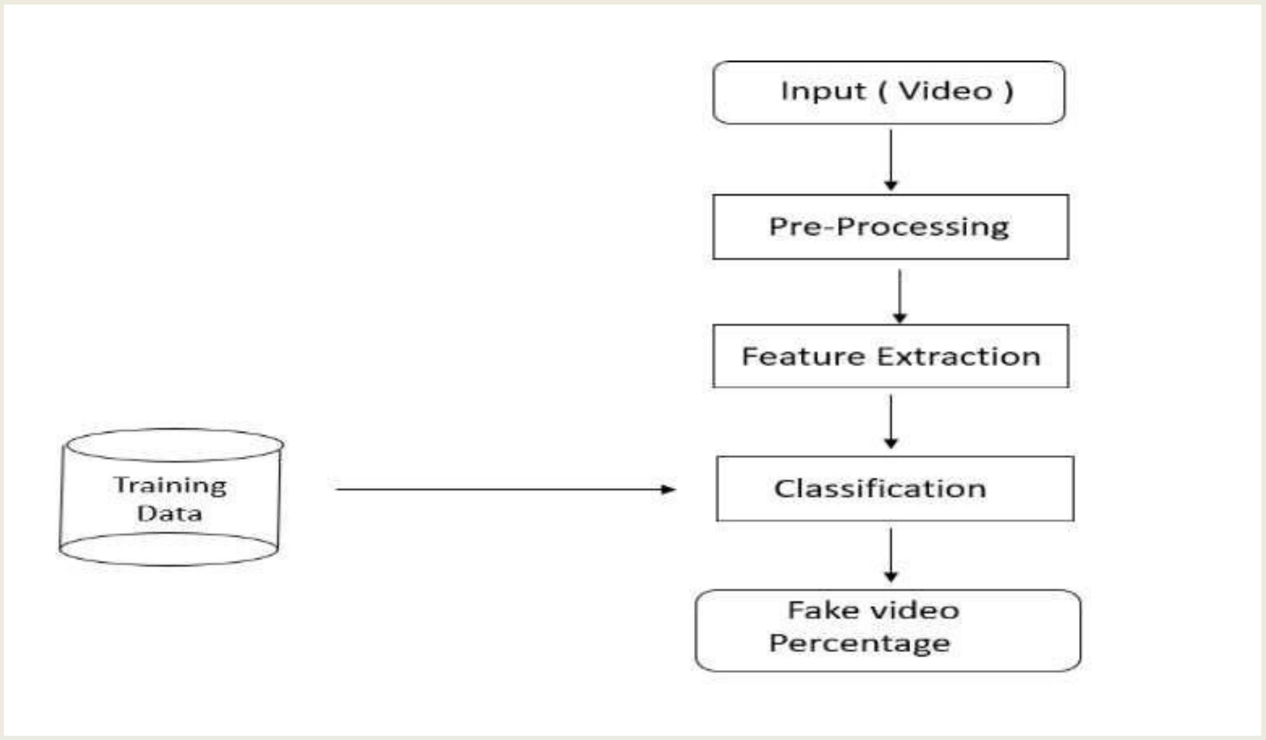


Figure 3.5.2.2: DFD level 1

3.5.2.3 DFD level 2

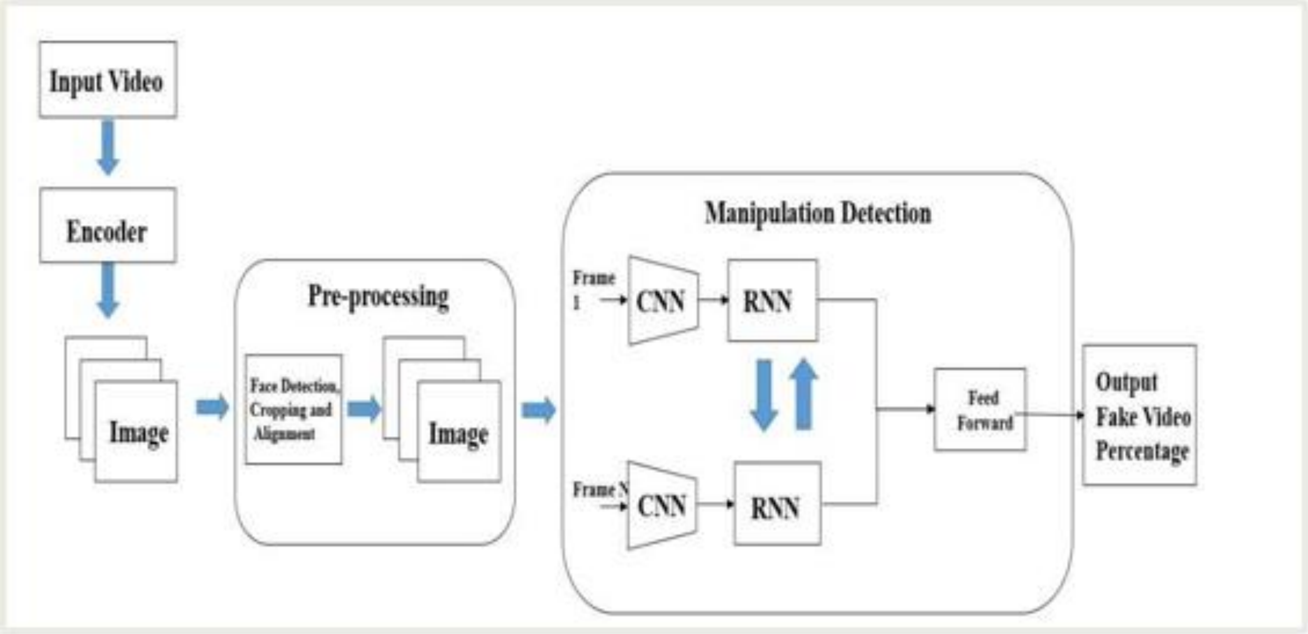


Figure 3.5.2.3: DFD level 2

3.5.3 Activity Diagrams

Training Workflow:

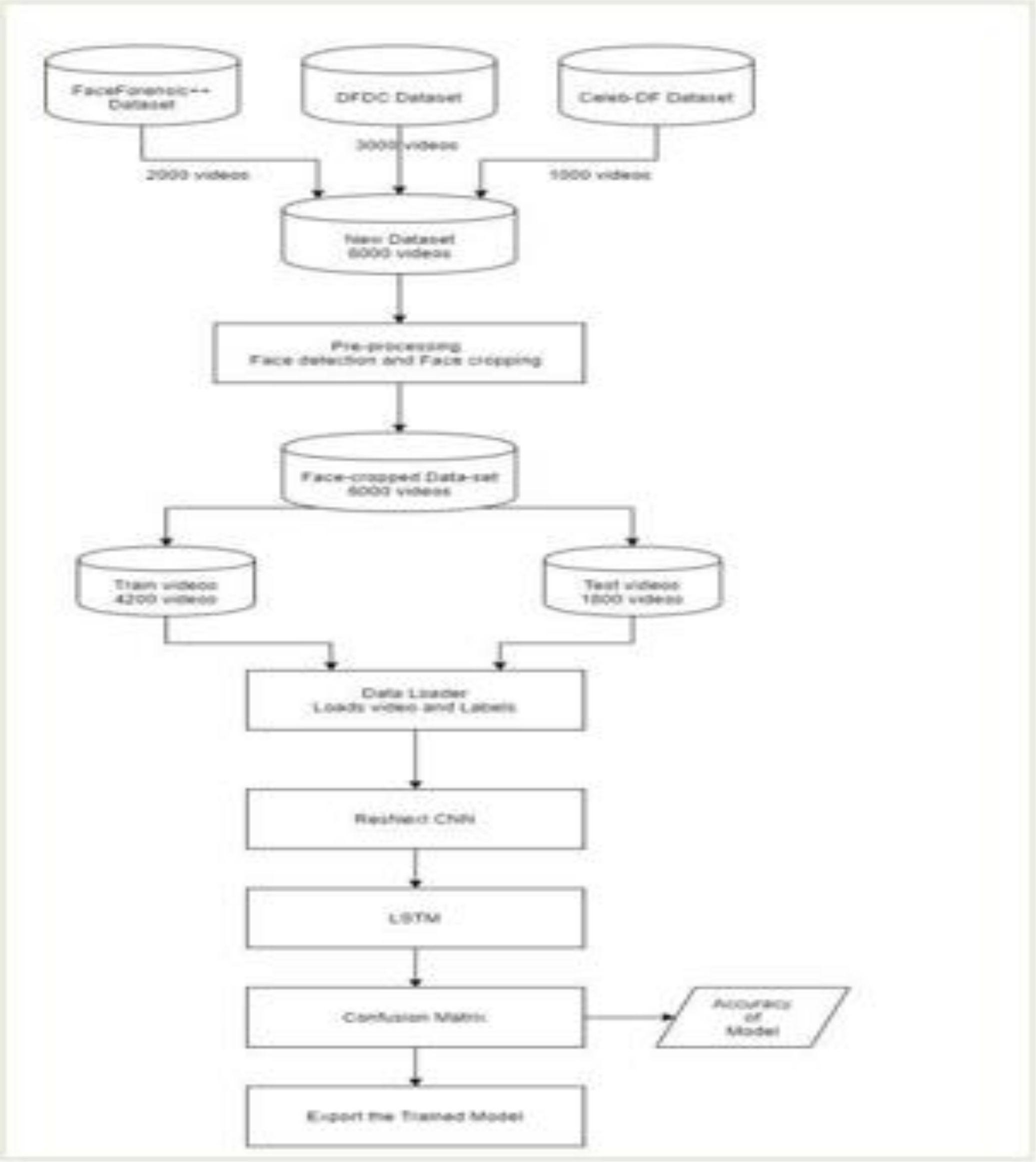


Figure 3.5.3.1: Training workflow

Testing workflow

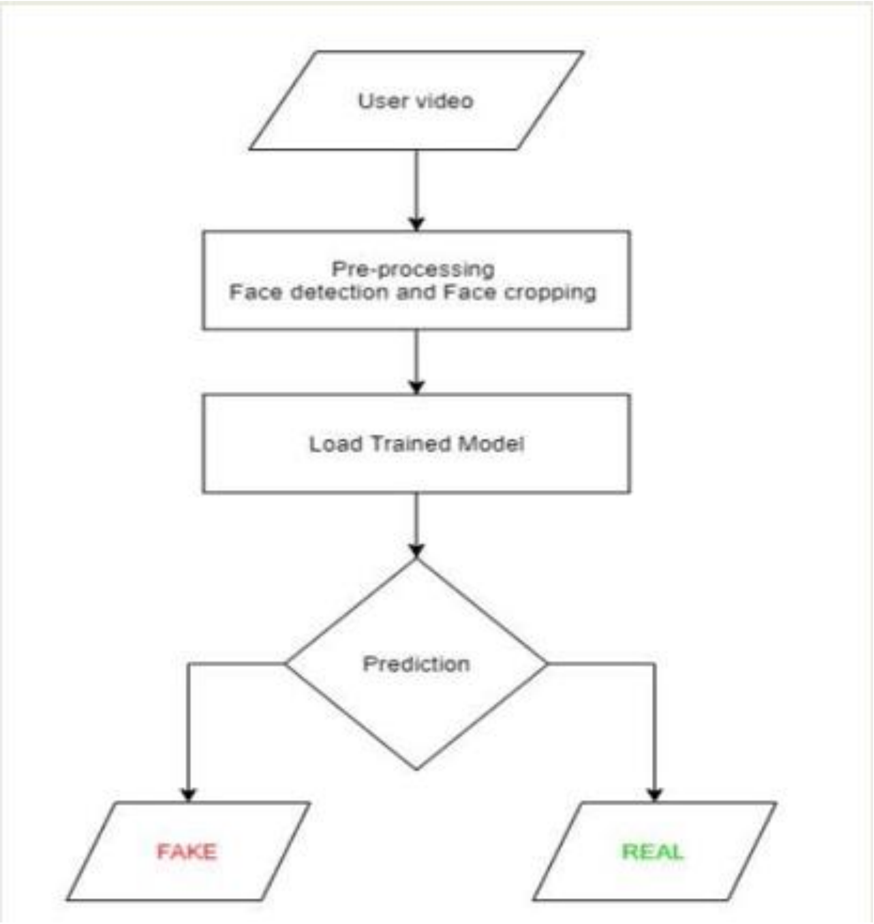


Figure 3.5.3.2: Testing workflow

3.5.4 Sequence Diagram

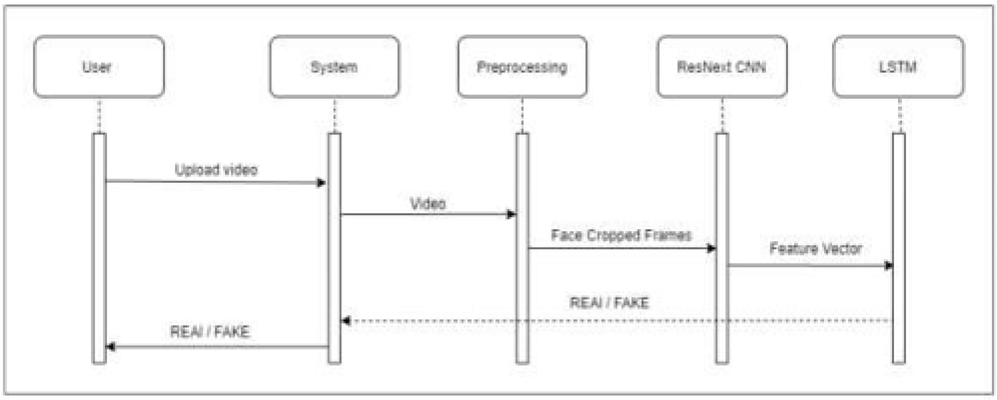


Figure 3.5.4.1: Sequence Diagram

3.6. Implementation plan/methodology

3.6.1. Entire Face Synthesis

Entire face synthesis aims to generate nonexistent fake face image xf from random vector v with neural network φ(·).

That is xf = φ(v).

For entire face synthesis tasks, GANs and VAEs are both feasible neural networks φ(·). However, according to the surveys (Verdoliva 2020; Nguyen et al. 2019d; Tolosana et al. 2020; Lyu 2020), GANs are the mainstream baseline technique. Many famous and popular entire face synthesis techniques such as PGGAN, StyleGAN, etc. are GAN-based and are able to generate high-quality DeepFake images. Compared with GANs, VAEs usually generate less realistic faces (i.e., being blurred). The reason why the images generated by VAEs tend to be blur is that the training principle makes VAEs assign a high probability to training data points, which cannot ensure that blurry data points are assigned to a low probability (Huang et al. 2018). Since the DeepFake images generated by VAE are not realistic enough, this section mainly introduces the GAN-related works. Using GANs for entire face synthesis is actually a kind of distribution mapping. The GANs learn the mapping from random distribution to human face distribution. Existing state-of-the-art methods can stably generate high-resolution images. which is benefited from the continuous improvement of the GAN network and training procedure. However, the current methods still suffer from the training difficulty (e.g., mode collapse problem of GAN training procedure). Furthermore, the generated images are not realistic enough due to the lack of general knowledge of face distribution (e.g., facial symmetry). As shown in the entire face synthesis part of Fig. 1, the fake images are very realistic and it is hard to distinguish real images from fake ones. Existing works mainly focus on improving the training stability, resolution, and controllable face attribute. The classical examples are deep convolutional GAN (DCGAN) (Radford et al. 2015), Wasserstein GAN (WGAN) (Arjovsky et al. 2017), progressive growing GAN (PGGAN) (Karras et al. 2017), and style-based GAN (StyleGAN) (Karras et al. 2019). The very first work which combines convolutional neural network (CNN) and GAN is a deep convolutional generative adversarial network (DCGAN) (Radford et al. 2015). It focuses on unsupervised learning and has comparable performance in image classification tasks with the pre-trained discriminator. The generator of it can easily manipulate lots of the semantic properties (i.e., manipulate attribute of a human face) of generated images profile from its interesting vector arithmetic properties. Two years later, there has been an explosion of in-depth research on GANs. Some GANs put emphasis on the stability of the GAN training. The groundbreaking work is Wasserstein-GAN (WGAN)

(Arjovsky et al. 2017). In the first published GANs, the procedure requires researchers to carefully maintain a balance between generator and discriminator. The mode dropping phenomenon also occurs frequently. To solve these hot potatoes, WGAN has theoretically minimized a reasonable and efficient approximation of the expectation-maximization (EM) distance, which only needs a few optimization designs on the original GANs. There are many types of research based on the WGAN. Gradient penalty WGAN (WGAN-GP) (Gulrajani et al. 2017) has indicated that WGAN sometimes still generates poor samples or fails to converge. The reason is that WGAN uses weight clipping to enforce a Lipschitz constraint. To improve the weight clipping operation, they have proposed to penalize the norm of the gradient of the discriminator with respect to its input fake image. The new designs train stably when generating high-quality home images. Simply using Wasserstein probability can not simultaneously satisfy sum invariance, scale sensitivity, and unbiased sample gradients. To improve it, Cramer GAN (CramerGAN) (Bellemare et al. 2017) has combined the best of the Wasserstein and Kullback–Leibler divergences to propose the Cramér distance. The CramerGAN performs significantly better than the WGAN. Boundary equilibrium GAN (BEGAN) (Berthelot et al. 2017) is also an improved version of WGAN (Arjovsky et al. 2017). To further balance the power of the discriminator against the generator, they have suggested pairing an equilibrium enforcing method with a loss derived from the Wasserstein distance together. They also have proposed a new way to control the trade-off between image diversity and visual quality. Some other works focus on how to generate highresolution images. The resolution of the images generate by them is at least 1024 × 1024. Meanwhile, the images are detailed, and it is quite difficult to distinguish between the genuine and the fake, which is very amazing. PGGAN (Karras et al. 2017) is the very first and famous work that proposes an effective method to generate high-resolution images. The resolution of the generated images is 1024×1024. It has proposed to progressively grow both the image resolution of the generator and discriminator. The images are starting from a low resolution and being detailed step by step with the new layers added in the model. This method is very reasonable in that it can speed up the training as well as greatly stabilize the GAN. However, the training procedure is still not good enough that some of the generated images are far from real. BigGAN (Brock et al. 2018) has attempted to generate highresolution diverse images from datasets such as ImageNet (Deng et al. 2009). They have applied orthogonal regularization to enforce the generator to be satisfied with a simple "truncation trick". Thus, the user can control the trade-off between image fidelity and variety by reducing the variance of the generator's input. To control the properties of generated images elaborately, StyleGAN (Karras et al. 2019) has proposed a new design to automatically learn the unsupervised separation of high-level attributes such as pose and human identity. The architecture also leads to stochastic variation in the generated images (e.g., freckles, hair). Furthermore, it enables intuitive, scalespecific control of the synthesis. StyleGAN2 (Karras et al.

2020) has exposed several typical artifacts of StyleGAN and has proposed changes in both model architecture and training methods to address them. In particular, they have encouraged good conditioning in the mapping from latent codes to images by the new design of generator normalization, progressive growing, and generator regularization. Different from the previous methods which use the GAN framework, generative flow (Glow) (Kingma and Dhariwal 2018) is a flow-based generative model that uses an invertible 1 × 1 convolution. The method is based on the theory that a generative model optimized towards the plain log-likelihood objective has the ability to generate efficient realistic-looking synthesis and manipulate large images.

3.6.2. Attribute Manipulation Definition

Attribute manipulation aims to modify facial properties P of a real face image xr to generate a new fake image xf with neural network φ(·, ·).

That is xf = φ(xr, P).

Using GANs for attribute manipulation is actually a kind of latent space editing. The key point is the quality of the GAN inversion technique. With a better attribute disentangle technique, the GANs for attribute manipulation can achieve more accurate attribute control. Existing state-of-the art methods [e.g., HifaFace (Gao et al. 2021d)] can perform accurate face editing while maintaining rich details of nonediting areas. However, the current methods are still limited by the labels in the training dataset. That is, it is difficult to control the attributes that do not exist in the label of the training dataset. As shown in the attribute manipulation part of Fig. 1, the real images are modified with facial attributes such as bald, blond hair, eyeglasses, etc. Existing works mainly focus on improving attribute manipulation accuracy. Attribute manipulation is also known as face editing, which can not only modify simple face attributes such as hair color, bald, smile, but also retouch complex attributes like gender, age, etc. The classical examples are StarGAN (Choi et al. 2018) and selective transfer GAN (STGAN) (He et al. 2019b). Invertible conditional GAN (IcGAN) (Perarnau et al. 2016) is the earliest attempt in GAN-based facial attribute manipulation. Based on an extension of the idea of conditional GAN (cGAN) (Mirza and Osindero 2014), they have evaluated encoders to map a real image into a latent space and a conditional representation, which allows the reconstruction and modification of arbitrary attributes of real human face images. The expression generative adversarial network (ExprGAN) (Ding et al. 2018) has added an expression controller module that can learn an expressive and compact expression code to the encoder-decoder network. The expression controller module enables it to edit photo-realistic facial expressions with controllable expression intensity. Previous studies can only perform image-to-image translation for two domains, which is cumbersome and timeconsuming. To be more efficient, StarGAN (Choi et al. 2018) has designed a single model to perform image-to-image translations for multiple domains. It allows simultaneous training of multiple different-domain datasets within a single network. As an improvement, StarGAN2 (Choi et al. 2020) simultaneously satisfies two properties in image-to-image translation: diversity of generated images as well as scalability over multiple domains. To represent diverse styles of a specific domain, they have replaced StarGAN's domain label with their domain-specific style code. To adapt the style code, they have proposed two modules: a mapping network and a style encoder. The style code can be extracted from a given reference image with a style encoder while the mapping network can transform random Gaussian noise into a style code. Utilizing these style codes, the generator learns to successfully synthesize diverse images over multiple domains. Although StarGAN is effective, due to the limitation of the content of the datasets, it can only generate a discrete number of expressions. To address this limitation, GAN animation (GANimation) (Pumarola et al. 2018) has introduced a novel GAN conditioning method based on action units (AU) annotations. It defines the human expression with a continuous manifold of the anatomical facial movements. The magnitude of activation of each AU can be controlled independently. Different AUs can also be combined with each other with this method. Most of the previous work inevitably changes the attribute irrelevant regions. To solve this problem, spatial attention GAN (SaGAN) (Zhang et al. 2018) propose a module to only change the attribute-specific region and keep the other area unchanged. This work properly exploits the attention mechanism to ensure a better face editing effect, which shows the feasibility of the attention mechanism in face manipulation. Previous methods have attempted to establish an attribute independent latent representation for further attribute editing. However, since the facial attributes are relevant, requesting for the invariance of the latent representation to the attributes is excessive. Therefore, simply forcing the attribute-independent constraint on the latent representation not only restricts its representation ability but also may result in information loss, which is harmful to attribute editing. To solve this problem, facial Attribute editing (AttGAN) (He et al. 2019b) has removed the strict attribute-independent constraint from the latent representation. It just applies the attribute classification constraint to the generated image to guarantee the correctness of attribute manipulation.Meanwhile, it groups attribute classification constraint, reconstruction learning, and adversarial learning together for high-quality facial attribute editing. The model supports direct attribute intensity control on multiple facial attribute editing within a single model. Considering that the specific editing task is only related to the changed attributes instead of all target attributes, as an improvement of AttGAN, STGAN (Liu et al. 2019) has selectively taken the difference between target and source attribute vectors as the input of the model. Furthermore, they have enhanced attribute editing by adding a selective transfer unit that can adaptively select and modify the encoder feature to the encoder-decoder. Mask-guided portraiting editing (MaskPE) (Gu et al. 2019) proposes a unique way to manipulate face attributes. They use a face parsing mask to guide the generation of face attributes. The main idea is to separately embed five facial components (i.e., left eye, right eye, mouth, skin & nose, and hair) into latent codes based on face parsing masks. Then they can modify any facial component independently. Due to the lack of paired images during training, previous methods typically use cycle consistency to keep the non-editing attributes unchanged. However, even if the cycle consistency is satisfied, images may still be blurry and lose rich details from input images for that the generator tends to find a tricky way (i.e., encodes the rich details of the input image into the output image in the form of hidden signals) to satisfy the constraint of cycle consistency. To solve this problem, Gao et al. (2021d) propose high-fidelity arbitrary face editing (HifaFace) to maintain rich details (e.g., wrinkles) of non-editing areas. Their work has two improvements. The first is that they directly feed the high-frequency information of the input image into the end of the generator with wavelet-based skip-connection, which relieves the pressure of the generator to synthesize rich details. The second is that they use another high-frequency discriminator as a complement to the image-level discriminator to encourage the image to have rich details.

3.6.3. Identity Swap Definition

Identity swap aims to replace the identity of source image xs by the identity ti of target image xt with neural network φ(·, ·) and generate a new fake image xf.

That is xf = φ(xs,ti).

As shown in the identity swap part of Fig. 1, the images in the fake videos have uneven qualities. Existing works mainly focus on improving the realism and resolution of the image. In general, the architectures used for these functions mainly fall into two categories: autoencoder-based and GAN-based. The classical works are cycle-consistent GAN (CycleGAN) (Zhu et al. 2017, 2021b). The methods which make the concept of DeepFake, especially identity swapping, become widely known are methods based on autoencoder. The autoencoder-based methods (OValery 2017) have no specific name or architecture. However, as they are all based on autoencoder, their pipeline is similar. The methods use one shared encoder and two independent decoders. The encoder and one of the decoders are trained by source identity while the encoder and the other decoder are trained by target identity. When the model is well trained, the encoder has the ability to extract the common features of source and target identities while the decoder records the specific features. At inference time, the image of the source identity goes through the encoder and the opposite decoder, producing a realistic swap. Nowadays, GAN-based methods are the mainstream in identity swap. The first work of the GAN-based method was CycleGAN (Zhu et al. 2017) proposed in 2017. In previous works, the absence of paired examples is always the limitation in image transformation tasks. CycleGAN has artfully solved this problem. Define a source domain X and a target domain Y , it builds a mapping G

: X → Y which is highly under-constrained and similarly constructs an inverse mapping F : Y → X. Then the cycle consistency loss which enforces F(G(X)) ≈ X (and vice versa) is the optimization target of the model. Through this circulation, there is no need for paired samples. Meanwhile, although not mentioned in the paper, the framework of CycleGAN can be used for identity swap easily. Faceswap-GAN (Lu 2018) is the implementation of CycleGAN which provides an identity swap functionality. It simply adds the adversarial loss and perceptual loss to encoder architecture. Face swapping GAN (FSGAN) (Nirkin et al. 2019) is a subject agnostic method that doesn't rely on the training of pairs of faces. It is also the first to simultaneously adjust the pose, expression, and identity variations for both a single image and a video sequence. The research in identity swap has been stagnated for a long time until the appearance of FaceShifter (Li et al. 2020b). It proposes a two-stage procedure for high fidelity and occlusion-aware face-swapping. Unlike many existing face-swapping works that leverage only limited information from the target image, FaceShifter generates the swapped face by thoroughly and adaptively exploiting the information of the target image. Appearance optimal transport (AOT) (Zhu et al. 2020) has formulated appearance mapping as an optimal transport problem. They have proposed an AOT model to formulate it in both latent and spatial space. In particular, a relighting module is designed to simulate the optimal transport plan. The optimization target is minimizing the Wasserstein distance of the learned features in the latent space, which enables better performance and less computation than conventional optimization. Information disentangling and swapping network (InfoSwap) (Gao et al. 2021a) aims to extract the most expressive information for identity representation. The main idea is to formulate the learning of disentangled representations as optimizing an information bottleneck trade-off. The information bottleneck principle provides a guarantee that in the latent space, areas scored as identity-irrelevant indeed contribute little information to predict identity. Megapixel level face swapping (MegaFS) (Zhu et al. 2021b) has proposed the first one-shot ultra-high-resolution face swapping method. To overcome the information loss in the encoder, they use a hierarchical representation face encoder (HieRFE) to find the complete face representation. Then they use a face transfer module (FTM) to control multiple attributes synchronously without explicit feature disentanglement. The contributions are ground-breaking. FaceInpainter (Li et al. 2021a) proposes a controllable face inpainting network under heterogeneous domains (i.e., oil painting, 3D cartoons, pencil drawing, exaggerated drawing, etc.). The framework has two stages. In the first stage, they use a styled face inpainting network (SFI-Net) to map the identity and attribute properties to the swapped face. The second stage contains a joint refinement network (JR-Net) that refines the attributes and identity details, generating occlusion-aware and high-resolution swapped faces with visually natural fused boundaries.

3.6.4. Expression Swap Definition

Expression swap aims to replace the expression of source image xs by the expression te of target image xt with neural network φ(·, ·) and generate a new fake image xf.

That is xf = φ(xs,te).

As shown in the expression swap part of Fig. 1, usually the mouth of the real images are changed. Existing works mainly focus on improving the diversity of input source and video consistency. Expression swap is also known as face reenactment. The classical examples are ICface (Tripathy et al. 2020) and SVGAN (Hyun et al. 2021). Face2Face (Thies et al. 2016) has proposed a threestep procedure. It first uses a global non-rigid model-based bundling approach to reconstruct the shape identity of the target human based on a prerecorded training sequence. Then it uses a transfer function to efficiently exploit deformation transfer in the low-dimensional semantic space. At last, the image-based mouth synthesis approach exploits the best matching mouth shapes offline sample sequence to generate a realistic mouth. A2V (Suwajanakorn et al. 2017) has used a recurrent neural network to train a model that can map from raw audio features of Obama's weekly address footage to mouth shapes. It is a cross-modal method that leverages the pronunciation features of the target person to synthesize the correct lip shapes for given audio content. It doesn't need an original video as expression-driven material. To match the input audio track, they have synthesized high-quality mouth texture and composited it with proper 3D pose matching to change what he appears to be saying. Pose-controllable audio-visual system (PC-AVS) (Zhou et al. 2021a) is another state-of-the-art cross-modal method. Previous audio-driven talking human face synthesis methods fail to model head pose, one of the key factors for talking faces to look natural. This is because pose information can rarely be inferred from audios. To solve this problem, PC-AVS introduces extra pose source video to compensate only for head motions and successfully disentangle the representations of talking human faces into the spaces of speech content, head pose, and identity respectively. Previous cross-modal methods only put emphasis on the lip motions and ignore the implicit ones such as head poses and eye blinks that have a weak correlation with the input audio. To model these implicit relationships, face implicit attribute learning generative adversarial network (FACIALGAN) (Zhang et al. 2021a) integrates the phonetics-aware, context-aware, and identity-aware information to synthesize the 3D face animation with realistic motions of lips, head poses, and eye blinks. Previous works may lose detailed information of the target leading to a defective output. To solve this problem, MarioNETte (Ha et al. 2020) has proposed a few-shot face reenactment framework that preserves the information of target identity even in situations where the facial characteristics of the source identity are far from the target. It has also introduced landmark transformation to cope with the varying facial characteristics of different people. Interpretable and controllable face reenactment network (ICface) (Tripathy et al. 2020) has proposed a two-stage neural network face animator which can control the pose and expressions of a given face image. The face animator is a data-driven and GAN-based system that is suitable for a large number of identities. Self-supervised video GAN (SVGAN) (Hyun et al. 2021) first puts emphasis on exploiting the discriminator of the GAN. They hypothesize two prominent constraints for realistic videos: consistency of appearance and coherency of motion. With these constraints, GANs are more likely to generate realistic videos. In other words, they have well defined what constraints should synthesized videos satisfy first. Wang et al. (2021) propose a one-shot neural talking-head synthesis approach. The method uses unsupervised learning to decompose for key features of an image: appearance feature, canonical keypoint, head pose, expression deformation. With the appearance feature and canonical keypoint of the source image, and synthesized with the head pose and expression deformation, a new fake image can be created. This work clearly disassembles the face information and reasonably exploits them. Most of the DeepFake detection methods did not take expression swap as the main detection objective. In our opinion, there are several reasons. As we can see from the previous description, expression swap has a similar technique to identity swap. Thus most of the detection methods are not specifically designed for them and only a few detection methods consider detecting expression swaps. On the other hand, it usually needs the coordination of audio to achieve a better display effect in the expression swap. Only the detection methods which simultaneously take images and audio into account are designed for this problem. The swapped expression strongly depends on the source video or image. The audio-video coordination opens the door for detection algorithm to tackle this problem from multiple angles, reducing the difficulty of this problem. Therefore, as we mainly investigate the DeepFake generation methods that are mentioned by the DeepFake detection methods, we take expression swap as an extension of identity swap in the survey.

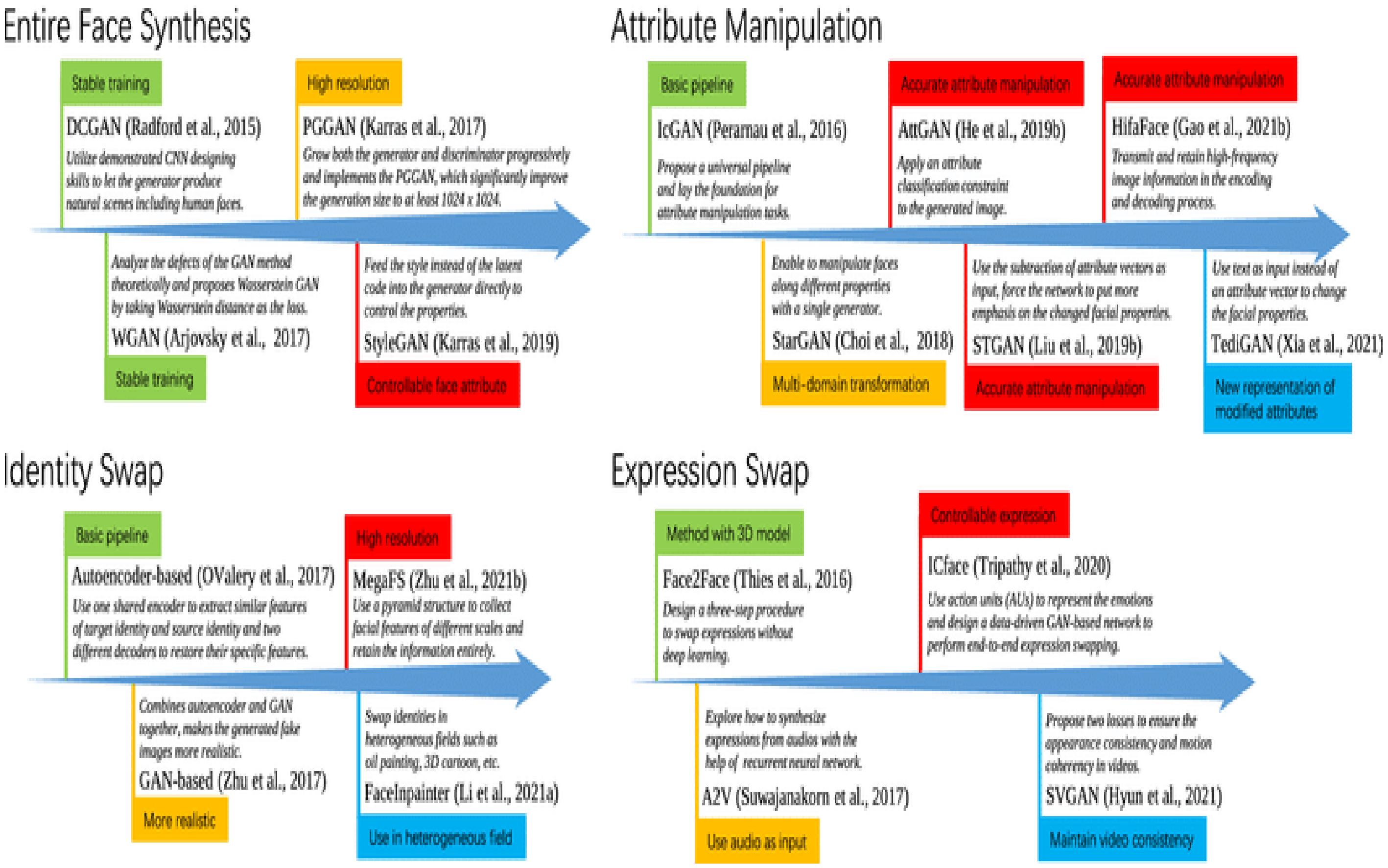

Entire Face Synthesis
Stable training
DCGAN (Radford et al., 2015)
Utilize demonstrated CNN designing skills to let the generator produce natural scenes including human faces.
High resolution
PGGAN (Karras et al., 2017)
Grow both the generator and discriminator progressively and implements the PGGAN, which significantly improve the generation size to at least 1024 x 1024.
Analyze the defects of the GAN method theoretically and proposes Wasserstein GAN by taking Wasserstein distance as the loss.
WGAN (Arjovsky et al., 2017)
Stable training
Feed the style instead of the latent code into the generator directly to control the properties.
StyleGAN (Karras et al., 2019)
Controllable face attribute
Attribute Manipulation
Basic pipeline
IcGAN (Peramau et al., 2016)
Propose a universal pipeline and lay the foundation for attribute manipulation tasks.
Accurate attribute manipulation
AttGAN (He et al., 2019b)
Apply an attribute classification constraint to the generated image.
Accurate attribute manipulation
HifaFace (Gao et al., 2021b)
Transmit and retain high-frequency image information in the encoding and decoding process.
Enable to manipulate faces along different properties with a single generator.
StarGAN (Choi et al., 2018)
Multi-domain transformation
Use the subtraction of attribute vectors as input, force the network to put more emphasis on the changed facial properties.
STGAN (Liu et al., 2019b)
Accurate attribute manipulation
Use text as input instead of an attribute vector to change the facial properties.
TediGAN (Xia et al., 2021)
New representation of modified attributes
Identity Swap
Basic pipeline
Autoencoder-based (OValery et al., 2017)
Use one shared encoder to extract similar features of target identity and source identity and two different decoders to restore their specific features.
High resolution
MegaFS (Zhu et al., 2021b)
Use a pyramid structure to collect facial features of different scales and retain the information entirely.
Combines autoencoder and GAN together, makes the generated fake images more realistic.
GAN-based (Zhu et al., 2017)
More realistic
Swap identities in heterogeneous fields such as oil painting, 3D cartoon, etc.
FaceInpainter (Li et al., 2021a)
Use in heterogeneous field
Expression Swap
Method with 3D model
Face2Face (Thies et al., 2016)
Design a three-step procedure to swap expressions without deep learning.
Controllable expression
ICface (Tripathy et al., 2020)
Use action units (AUs) to represent the emotions and design a data-driven GAN-based network to perform end-to-end expression swapping.
Explore how to synthesize expressions from audios with the help of recurrent neural network.
A2V (Suwajanakorn et al., 2017)
Use audio as input
Propose two losses to ensure the appearance consistency and motion coherency in videos.
SVGAN (Hyun et al., 2021)
Maintain video consistency

Figure 3.6.1.: Methodologies

CHAPTER 4

RESULTS ANALYSIS AND VALIDATION

4.1. Implementation of solution

4.1.1. Analysis:

Based on the categories of Deepfake detection 21 studies use special artifacts-based features generated by various editing processes. Among them, 20 studies use texture and Spatio-temporal consistent features, 14 studies involve facial landmarks-based features. Also, 13 research papers perform experiments using visual artifacts-based elements, for example, eye blinking, head posing, lip movement, etc. Eight pieces of work apply biological characteristics, whereas seven studies concern intra-frame inconsistencies with frequency domain analysis. In addition, six studies use GAN based features, and four studies cover latent space-based features.

4.1.1.1. Machine Learning based methods:

Machine learning based creates a feature vector by defining the right features using various state-of-art feature selection algorithms. It then feeds this vector as input to train a classifier to classify whether the videos or images are manipulated by Deepfake or not.

GANs are used to automatically train a generative model by treating the unsupervised issue as supervised and creating photo-realistic fake faces in images or videos. Some ML-based methods aspire to show certain irregularities found in such GANs generated fake videos or images.

To fool the users, most techniques modify certain regions of the face, such as shade of the eyes, ear with a ring, etc. Such methods using a single part (a.k.a. feature) are limited to identifying or detecting the manipulated area.

As far as the performance concern in machine learning based Deepfake methods, it is observed that these approaches can achieve up to 98%

accuracy in detecting Deepfakes. However, the performance entirely relies on the type of dataset, the selected features, and the alignment between the train and test sets. The study can obtain a higher result when the experiment uses a similar dataset by splitting it into a certain level of ratio, for example, 80% for a train set and 20% for a test set. The unrelated dataset drops the performance close to 50%, which is an arbitrary assumption.

4.1.1.2. Deep Learning based methods:

Deep learning models have been used widely due to their feature extraction and selection mechanism, as they can directly extract or learn features from the data. GAN simulator that replicates collective GAN-image artifacts and feeds them as input to a classifier to identify them as Deepfake. Deep learning-based method was proposed in for Deepfake video detection. Two inception modules, (i) Meso- 4 and (ii) MesoInception-4, were used to build their proposed network. In this technique, the mean squared error (MSE) between the actual and expected labels is used as the loss function for training. Further innovations are achieved by using an architecture named capsule-network (CN). The CN needs a smaller number of parameters to train than very deep networks. An ensemble learning technique is applied to increase such structures' performance, which achieves more than 99% accuracy.

4.1.1.3. Statistical Measurements based methods:

Statistical measurement based on the use of the information-theoretic study for validation. In these models, the shortest paths are calculated between original and Deepfake videos/images distributions.Different statistical measures such as average normalized cross-correlation scores between original and suspected data helps to understand the originality of the data examined the photo response non uniformity (PRNU) for detecting Deepfakes in video frames. PRNU is a unique noise pattern in the digital images that occurred due to the defects in the camera’s light-sensitive sensors. Because of its distinctiveness, it is also considered the fingerprint of digital photos. The research generates a sequence of frames from input videos and stores them in chronologically categorized directories. Each video frame is clipped with the same pixel range to preserve and clarify the portion of the PRNU sequence. These frames are then divided into eight equal groups. After that, it correlates them by measuring the normalized cross-correlation scores and calculating the differences between the correlation scores and the mean correlation score for each frame.

To conduct a granular analysis, the research adopts a temporal perspective, generating sequences of frames from input videos. These frames are meticulously organized into chronologically categorized directories, facilitating a systematic examination of the temporal evolution of content. This temporal analysis is crucial for detecting subtle variations or anomalies that may emerge over the course of a video, adding an additional layer of depth to the validation process.

The frames, having undergone pixel range clipping, are stratified into eight equal groups. This segmentation facilitates a structured correlation analysis, where the normalized cross-correlation scores are measured and differences from the mean correlation score are calculated for each frame. This approach allows for a fine-grained examination, pinpointing deviations in correlation patterns that may signify potential manipulation. The stratified correlation analysis enhances the sensitivity of the models to subtle alterations in content, further fortifying the validity of the detection process. This organizational strategy not only aids in systematic analysis but also ensures that the models are equipped to detect temporal inconsistencies or anomalies that may be indicative of manipulation. The meticulous chronological organization enhances the models' ability to discern nuanced patterns within the temporal evolution of content.

Category	Accuracy %
Deep Learning based methods	99%
Machine learning based methods	98%
Statistical based methods	

Table 4.1.1.3.1. Accuracy based on model

4.1.2 Design drawings/Schematics/Solid models:

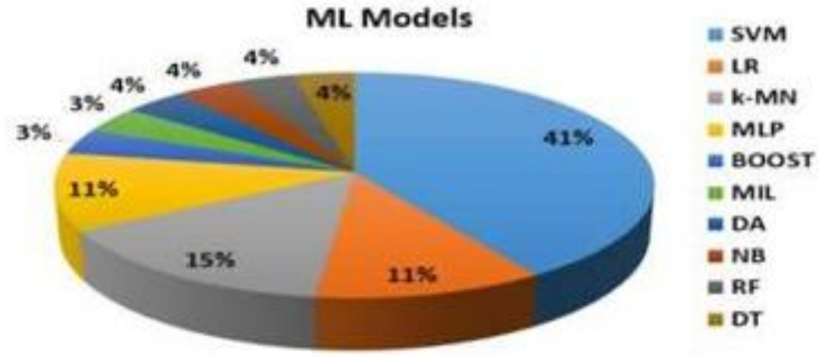


Fig.4.1.2.1. Machine Learning Model

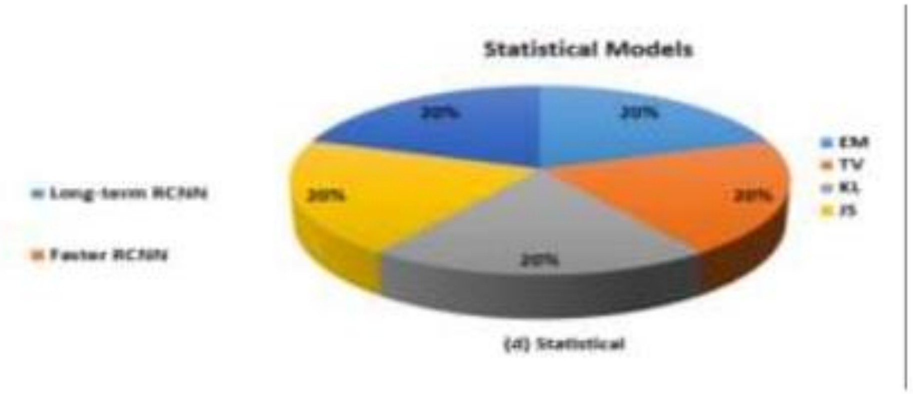


Fig.4.1.2.2. Statistical Learning Model

4.1.3. Report Preparation:

This segment attempts to decide the efficacy of Deepfake detection methods. The output assessment values are first obtained and stored in an Excel document based on the studies. After that, we count the number of studies that use the same method and the same measurement metrics (precision, accuracy, and recall).

And finally, we apply four metrics: the minimum, maximum, mean, and standard deviation, correlation, covariance, on these values based on the mean values of accuracy and deep learning-based methods outperform other methods and achieve 89.73% and 0.917 respectively. Based on the overall results, we found deep learning-based techniques are efficient for detecting Deepfake.

4.2. Testing/Characterization/ Interpretation/ Data validation

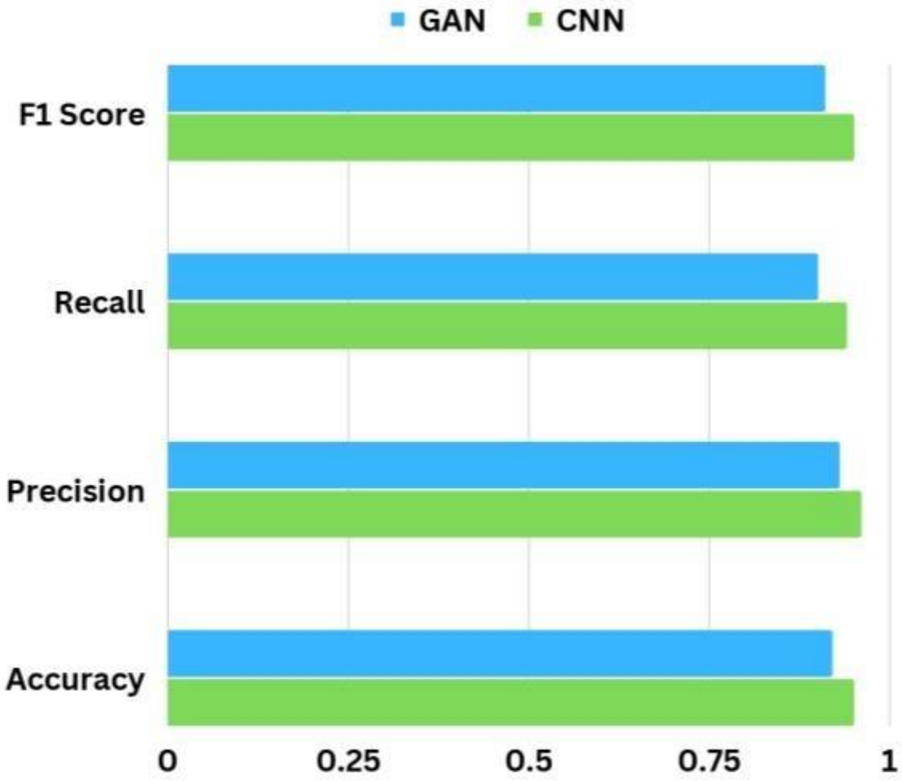
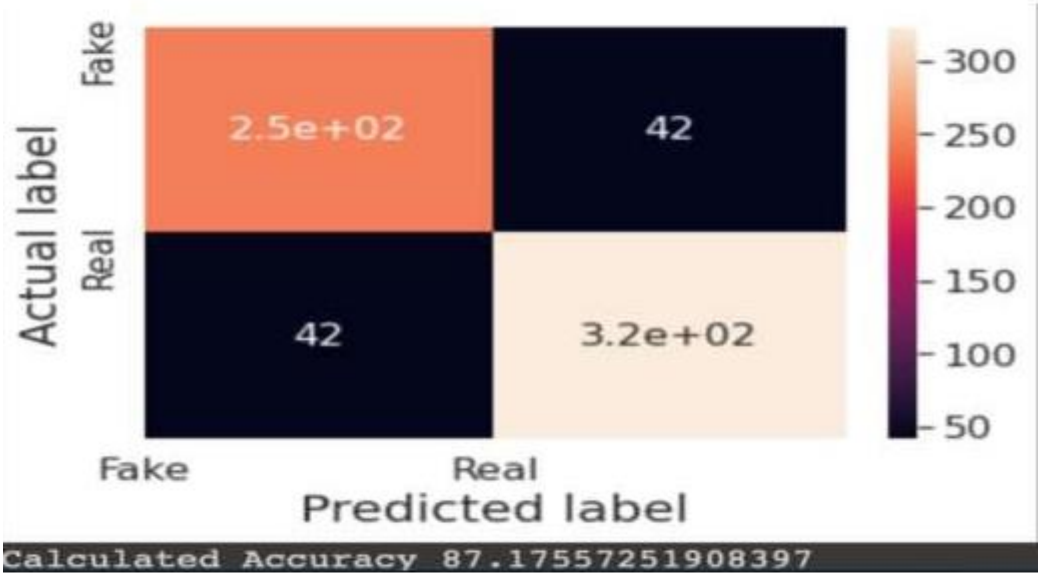


Figure 4.2.1: Result accuracy of different algorithms

Model name	Accuracy	Precision	Recall	F1 score
CNN	0.95	0.96	0.94	0.95
GAN	0.92	0.93	0.90	0.91

Confusion Matrix:

The number of correct and incorrect predictions are summarized with count values and broken down by each class the matrix formed is called confusion matric which gives summary of prediction results on a classification problem. The confusion matrix shows the ways in which our classification model might get confused when it makes predictions. It gives us insight on the errors and type of errors being made. It is used to evaluate our model and calculate the accuracy.



```
Calculated Accuracy 87.17557251908397
```

Figure 4.2.2: Confusion matrix

Where value of e = 2.71828182……

Type of Testing

Used Functional

Testing

1. Unit Testing
2. Integration Testing
3. System Testing
4. Interface Testing

Non-functional Testing

1. Performance Testing
2. Load Testing

3. Compatibility Testing

Table 4.2.1. Test Case report

Case id	Test Case Description	Expected Result	Actual Result	Status
1	Upload a word file instead of video	Error message: Only video files allowed	Error message: Only video files allowed	Pass
2	Upload a 200MB video file	Error message: Max limit 100MB	Error message: Max limit 100MB	Pass
3	Upload a file without any faces	Error message:No faces detected. Cannot process the video.	Error message:No faces detected. Cannot process the video.	Pass
4	Videos with many faces	Fake / Real	Fake	Pass
5	Deepfake video	Fake	Fake	Pass
6	Enter /predict in URL	Redirect to /upload	Redirect to /upload	Pass
7	Press upload button without selecting video	Alert message: Please select video	Alert message: Please select video	Pass
8	Upload a Real video	Real	Real	Pass
9	Upload a face cropped real video	Real	Real	Pass
10	Upload a face cropped fake video	Fake	Fake	Pass

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1. Conclusion:

We present basic techniques and discuss different detection models' efficacy in this work. We summarize the overall study as follows:

- We proposed a generalized detection method to detect three types of deepfake techniques: face swap, lip-syncing and puppet-master.
- Faceswap uses Encoder- decoder, GAN, and CNN technique for deepfake learning.
- Lip- Syncing Generation uses RNN, Encoder–decoder CNN, Temporal GAN.
- Puppet- master Generation uses mainly GAN for deepfake learning.
- We found three types of common traces (residual noise, warping artifacts, and blur effects) generated by the deepfake process and applied them to the proposed network for deepfake detection.
- They tried to do steganalysis for the base network to detect residual noise.
- Landmark patches were extracted from the semantic facial region to detect warping artifacts, which are unnatural high-level features.
- Then finally applying IQM features effective identification of blurring effects.
- The deep learning-based methods are widely used in detecting Deepfake.
- The deep learning models hold a significant percentage accuracy in all the models.
- The most widely used performance metric is detection accuracy.
- Deep learning techniques are effective in detecting Deepfake.
- Results revealed that each detection strategy is effective, and the performance of the proposed network is superior to that of existing networks.

5.2. Future work:

After analyzation of various approaches one of the shortcomings is of the memory space for storing huge number of videos for the training purpose of the model and also for computing the result more efficiency needs to be integrated with the machine learning model. This can be thought of next challenge for deepfake detection and sorting them out.

In this meta world, more and more videos need to be generated and for storage purpose concept of virtual memory could play an important role which depicts as if RAM has huge space while it's not.

Except for the images and videos detection whether it is fake or not, audios also need to be identified. In audio files detection of machine generated voice and mis-match in pitch or modulation could be identified by building different machine model and then training and testing can be done for achieving higher accuracy and efficiency.

Addressing Memory Space Challenges in Deepfake Detection:

A critical observation in various approaches highlights a significant challenge related to the memory space required for storing an extensive number of videos during the training phase of the model. To overcome this limitation, there is a pressing need to integrate more efficient memory management strategies into the machine learning model. The exploration of innovative solutions, such as optimized data compression techniques or leveraging distributed computing frameworks, could be instrumental in mitigating the memory space constraints and ensuring the scalability of deepfake detection systems.

Virtual Memory in the Meta World:

In the evolving meta world, where the generation and storage of videos are becoming increasingly prevalent, the concept of virtual memory emerges as a potential solution. Virtual memory, with its ability to create an illusion of extensive RAM space while efficiently utilizing storage resources, presents itself as a promising avenue. Integrating virtual memory concepts into deepfake detection systems could optimize the utilization of computational resources, providing an effective workaround for the inherent challenges associated with limited physical memory.

Expanding Scope to Audio File Detection:

Beyond the realm of visual content, the comprehensive nature of deepfake detection demands an extension to audio file identification. Recognizing the importance of audio in the context of manipulation, a dedicated effort is required to build machine models specifically tailored for audio detection. This entails developing models that can discern machine-generated voices, detect discrepancies in pitch or modulation, and effectively identify manipulated audio content. The integration of advanced machine learning algorithms for audio analysis opens up new dimensions in achieving higher accuracy and efficiency in overall deepfake detection.

Multi-Modal Approach for Holistic Detection:

A holistic approach to deepfake detection involves embracing a multi-modal perspective that encompasses images, videos, and audios. By concurrently training and testing different machine models for each modality, a comprehensive understanding of the authenticity of multimedia content can be achieved. This multi-modal approach not only enhances the accuracy of detection but also ensures a more robust defense against increasingly sophisticated deepfake techniques that may span multiple modalities.

Continuous Model Optimization and Adaptation:

The dynamic landscape of deepfake technology underscores the importance of continuous model optimization and adaptation. As new challenges, such as memory space constraints and multi-modal detection, are addressed, it is essential to implement strategies for ongoing refinement. This includes staying abreast of advancements in virtual memory technologies, exploring novel audio detection algorithms, and incorporating these insights into the training and testing processes. Continuous adaptation ensures that deepfake detection systems remain resilient in the face of emerging complexities.

Collaborative Research for Collective Progress:

Recognizing that deepfake detection is a multifaceted challenge, collaboration across the research community becomes paramount. By fostering collaborative efforts, researchers can collectively pool expertise, share insights, and collectively address the diverse challenges associated with memory space constraints, virtual memory integration, and multi-modal detection. This collaborative approach accelerates progress, promoting the development of more robust and versatile deepfake detection systems.

In summary, overcoming memory space challenges, integrating virtual memory concepts, expanding into audio file detection, adopting a multi-modal approach, ensuring continuous model optimization, and fostering collaborative research are integral components of addressing the evolving complexities in deepfake detection within the meta world.

REFERENCES

1. Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. arXiv preprint arXiv:1809.00888, 2018.

2. Belhassen Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In ACM Workshop on Information Hiding and Multimedia Security, pages 5–10, 201

3. Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In ACM Workshop on Information Hiding and Multimedia Security, pages 1–6, 2017.

4. Francois Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In IEEE Conference on Computer Vision and Pattern Recognition, 2017

5. Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In IEE Workshop on Information Forensics and Security.

6. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning.

7. Johnson , Evelyn. "Prepare for a Long Battle Against Deepfakes - KDnuggets." KDnuggets, Feb. 2020, www.kdnuggets.com/2020/02/long-battle-against-deepfakes.html.

8. Bitchkei, Silvia. "Deepfakes and Cybercrime: An Introduction - Hitachi Systems Security." Systèmes De Sécurité Hitachi, 14 Nov. 2019, hitachi-systems-security.com/deepfakes-and- cybercrime-an-introduction.