International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

Investigating the Evolving Landscape of Deepfake Technology: Generative AI's Role in it's Generation and Detection

Mrs Supriya Shree[1], Riddhi Arya[2], Saket Kumar Roy[3]*
[1]Assistant Professor, Department Of Computer Science,St. Xavier's College Of Management & Technology,Patna,Bihar,India.
[2,3]UG – Bachelor's in Computer Application,St. Xavier's College Of Management & Technology,Patna,Bihar,India
***Emails**: supriya@sxcpatna.edu.in[1], aryariddhi03@gmail.com[2], saketkroy06@gmail.com[3]*
****Orchid ID:** https://orcid.org/0009-0003-0760-0922*

Abstract
The world of artificial intelligence is constantly changing, with Generative AI and Large Language Models (LLMs) leading the way in bringing new technological advancements. This paper offers a detailed look at these groundbreaking technologies and how they are shaping the digital world today. We explore the technical aspects of Generative AI and LLMs, explain their unique features, and compare them to traditional AI models.One of the key focuses of our research is the growing issue of DeepFakes—artificial intelligence-generated media that presents a significant challenge in verifying content. We conduct a thorough examination of few deepfake detection techniques out of which we will be implementing and analyzing one of them. Our research implements a framework for Deep Fake Image Detection. The suggested solution utilizes a RESNET-50(Residual Network with 50 layers) and MTCNN (Multi-task Cascaded Convolutional Networks) models for detecting whether the images are real or fake. This study conducts the Hypothesis testing for the proposed solution taking in consideration that the current Deepfake detection algorithms are less effective in detecting highly realistic Deepfakes compared to less sophisticated manipulations. By investigating the convergence of deep learning, neural networks, and sophisticated algorithms, we set the stage for advancements in AI-based content verification.
***Keywords:** AI-Based Content Verification; Artificial Intelligence; Deep Fake Image Detection; Deepfakes; Generative AI.*

1. Introduction

Recent developments in generative AI, catalyzed by ChatGPT, have become a focal point of discussion. Generative AI possesses the potential to contribute across a myriad of domains, encompassing natural language generation, translation, and the generation of diverse and imaginative content. Of particular significance, Large Language Models (LLM) such as OpenAI's GPT series have demonstrated groundbreaking outcomes in the realm of natural language comprehension and generation. The wave of generative AI, instigated by ChatGPT, has extended its influence to encompass visual mediums, including Stable Diffusion and Midjourney, capturing the attention of the general populace. Considering these advancements, LLM are being extensively harnessed across diverse domains. Implementing extensive training on enormous volumes of textual data, LLM exhibit the ability to comprehend and generate natural language. Building upon this prowess, LLMs find application in various domains, including customer interactions, creative content creation, and question-answering. In the realm of Generative AI, ANNs serve as the foundation for creating models that can generate new content, while LLMs like GPT-4 are advanced instances of Generative AI, leveraging vast neural networks to

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

produce text that's remarkably human-like in its coherence and relevance. [5][6]

1.1. AI, A Boon or a Curse?

The notable advances in artifcial neural network (ANN) based technologies play an essential role in tampering with multimedia content. For example, AI-enabled software tools like FaceApp , and FakeApp have been used for realistic-looking face swapping in images and videos. This swapping mechanism allows anyone to alter the front look, hairstyle, gender, age, and other personal attributes. The propagation of these fake videos causes many anxieties and has become famous under the hood, Deepfake. This research comprehensively examines the dual aspects of Generative AI and Large Language Models (LLMs). On one hand, it highlights the remarkable advancements and conveniences that AI has brought to our lives, transforming complex tasks into simple ones. On the other hand, it delves into the potential perils—how, if misused, this technology could inflict significant harm on individuals' lives and reputations. An integral part of our study is dedicated to exploring effective strategies to mitigate the burgeoning threat of deepfakes. By harnessing the very technologies that enable their creation, namely Generative AI and LLMs, we aim to develop robust countermeasures to safeguard against the misuse of these powerful tools.

Figure 1 Global Statistics [7]

The global market for creative AI technology is expected to grow to USD 76.8 billion by 2030. This growth, at a rate of 31.5% every year, shows how important and in-demand this technology is. The reasons for this big growth include new and better AI methods, its use in many areas like video games and healthcare, helping businesses do things faster and cheaper, and making things that are special for each person in areas like shopping and customer help. However, there are worries about how this technology might be used in the wrong way, like making fake videos that seem real. So, as this technology gets better, there will be new rules to make sure it's used in a good way in Figure 1.[2]

1.2. Technical Aspects of Generative AI & Llms

Generative AI makes fresh, real-looking stuff by studying old data. It uses smart code, like neural networks and machine learning, to produce- new images, text, music, and more. The key to generative AI is foundation models like GPT (generative pre-trained transformers). These models train on huge datasets to do various tasks with some extra fine tuning Figure 2. Large Language Models (LLMs) are part of generative AI. They focus on understanding and creating text like humans. LLMs work with the transformer architecture, which is good at handling text data. LLMs train on massive amounts of text, allowing them to grasp complex language patterns.[2]

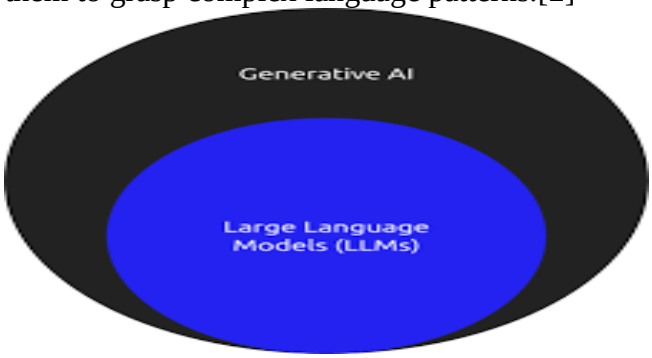

Figure 2 Venn Diagram [8]

1.3. Key Features of Generative AI & Llms

1.3.1. Generative AI

Learning from Data: It can learn from existing data to generate new content that reflects the characteristics of the training data without repeating it.

Diverse Content Creation: Capable of producing a wide range of content, including

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

images, videos, music, speech, text, software code, and product designs.

Foundation Models: Uses AI foundation models like generative pre-trained transformers (GPT) for various tasks with additional fine-tuning.

Business Applications: Enhances customer experience, increases employee productivity, and accelerates product development.

1.3.2. Large Language Models

Understanding Language: Designed to understand, generate, and manipulate human language with a high level of sophistication.

Transformer Architecture: Operates on a transformer-based architecture, effective in handling sequential data like text.

Versatile Applications: Can be used for language translation, text summarization, sentiment analysis, and creative writing.

Contextual Understanding: Capable of generating coherent and contextually relevant text, making them useful for content creation and automated customer support. As show in Table 1.

Table 1 Traditional AI and Modern AI Model

Traditional AI Model	Modern AI Model
It relies on logic programmed by humans.	Modern AI learns from data using ML Algorithms.
These are static and don't learn over time.	Modern AI System can continuously learn over time and improve performance.
These are designed for specific tasks within a narrow domain.	Modern AI can handle a variety of tasks of different domains.
They struggle with unstructured data such as images and speeches.	Modern AI excels in learning and analyzing from such kind of data.
Traditional AI doesn't have excellent computational Powers	Modern AI is equipped with excellent computational Powers with an exponential growth in it.

1.4. Applications of Generative AI and LLMs

Healthcare: Generative AI and LLMs are revolutionizing healthcare by aiding in drug discovery, personalizing patient care plans, and providing virtual assistance for routine inquiries. They help analyze large datasets to predict patient outcomes and assist in diagnosing diseases from medical images.

Automotive: In the automotive sector, these technologies are driving advancements in autonomous vehicles, enhancing safety features, and personalizing the in-car experience. They contribute to the design process by generating models for new vehicle parts and optimizing supply chain logistics.

Finance: Financial institutions employ Generative AI and LLMs for fraud detection, risk management, and automated customer service. They can simulate market scenarios to inform investment strategies and generate financial reports, making complex data more accessible.

Content Verification: Generative AI aids in content verification by detecting deepfakes, plagiarism, and ensuring the authenticity of digital media. LLMs can cross-reference information against vast databases to validate facts and flag inconsistencies.

Others: Beyond these fields, Generative AI and LLMs find applications in education for personalized learning, in legal for document analysis, in gaming for creating dynamic narratives, and in retail for enhancing customer experience through chatbots and product recommendations. [1][2][3].

2. Content Verification

Imagine: You are browsing the web and come across an article that claims to reveal a shocking truth about a political leader. The article is accompanied by a photo that shows the leader in a compromising situation. You are intrigued and outraged by the story, but you also wonder: is this real or fake? The domain of content verification within Generative AI and LLMs grapples with the challenge of discerning AI-created content from

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

human-made, a task made difficult by the rapid advancement and sophistication of generative technologies. Key challenges include the potential for AI to propagate misinformation due to inaccuracies or fabrications, the reflection of biases present in training data, and the phenomenon of "AI hallucinations" where AI fabricates plausible but false information. These issues underscore the need for robust verification techniques such as cross-referencing with credible sources, employing fact-checking tools, and metadata analysis. As AI-generated content becomes more prevalent, the combination of technical and manual methods is crucial for maintaining information integrity and countering misuse, with the field evolving to develop more sophisticated AI models, enhance human-AI collaboration, and implement stronger watermarking and provenance tracking to ensure content traceability in Figure 3.

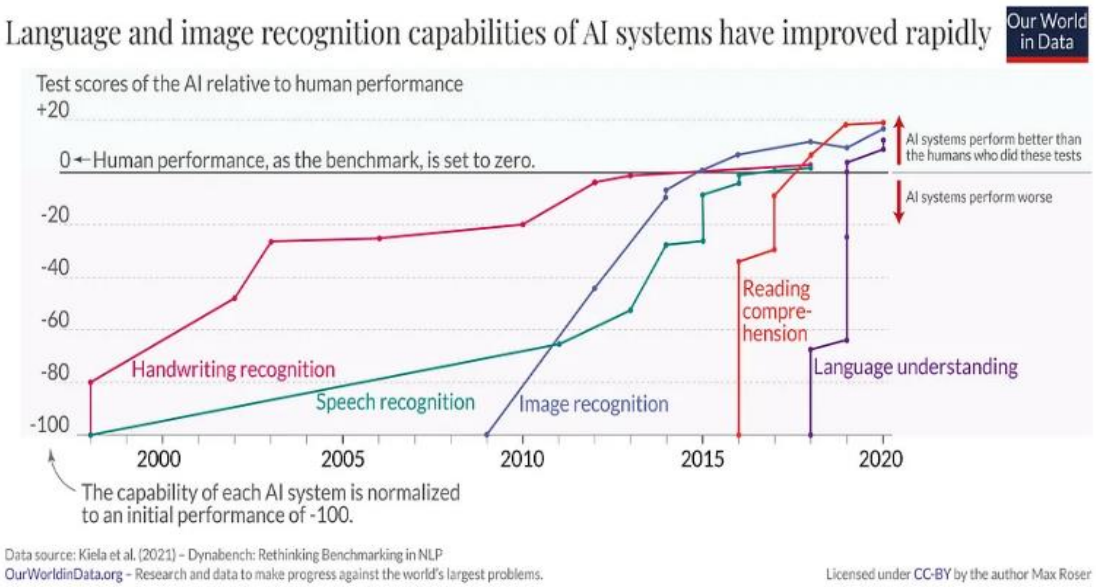

Figure 3 Lang & Image Recognition by AI [9]

The chart below illustrates the rapid evolution of AI systems in the past two decades, showcasing their remarkable progress in language and image recognition. Starting from an initial performance of -100, AI systems have advanced to consistently outperform humans in various domains, marking a significant shift from a decade ago when such feats were inconceivable.[5]

2.1. Applications of Content Verification

Social Media Moderation: AI helps social media platforms identify inaccuracies, false information, and harmful content like hate speech, enhancing user safety.

Fact-Checking: AI accelerates the fact-checking process by sifting through vast data to identify inconsistencies and errors, allowing for real-time verification of news stories.

Deepfake Detection: AI analyzes videos and images to detect deepfakes by identifying inconsistencies that indicate manipulation, thus preventing the spread of false narratives.

Automated Journalism: AI algorithms generate news stories without human intervention, covering a wider range of topics quickly and accurately.

Information Verification: AI transforms the verification process by analyzing content against trusted sources using natural language processing (NLP) algorithms.

Fraud Detection: In finance, AI and data

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

analytics significantly impact risk assessment and fraud detection, analyzing patterns to identify fraudulent activities.

Healthcare: Precision medicine and patient treatment are enhanced through big data analysis of electronic health records (EHRs), leading to better healthcare outcomes.

Manufacturing and Supply Chains: AI-driven predictive maintenance reduces operating expenses and equipment downtime, optimizing manufacturing processes.

3. Deepfake Detection & Analysis

In 1860, there was the first instance of manipulated multimedia content, where a portrait of John Calhoun was altered to Feature Abraham Lincoln's head. The initial deepfake emerged in September 2017, when a Reddit user known as "deepfake" shared videos of actresses with their faces swapped onto explicit material. Another well-known case involved the deepNude app, enabling users to produce fake nude images. This marked the moment when deepfakes started gaining significant attention and popularity among the public.Today, it is very easy for anyone to create fake videos using deepfake technology like FakeApp, FaceSwap, and ZAO without needing a background in computer engineering. Additionally, there are open-source projects like DeepFaceLab on GitHub and tutorials available on YouTube. In 2018, a deepfake video of former U.S. President Barak Obama went viral online, where he appeared to insult then-President Donald Trump. In June 2019, a fake video of Facebook CEO Mark Zuckerberg was shared on Instagram by the Israeli advertising company "Canny". Furthermore, there have been extremely realistic deepfake videos of Tom Cruise on TikTok that quickly gained 1.4 million views in just a few days. In addition to visual manipulation, audio deepfakes pose a new threat in the realm of cyber-attacks. These deceptive audio techniques, such as WaveNet, Tacotron, and deep voice, have the potential to cause serious harm to individuals. The prevalence of fake audio-assisted financial scams has surged in 2019 as a result of advancements in speech synthesis technology in Figure 4. For instance, in August 2019, a CEO of a European company fell victim to an audio deepfake and mistakenly transferred $243,000. The perpetrator utilized a voice-mimicking AI software to replicate the CEO's speech patterns by training machine learning algorithms with audio samples sourced from the internet. "If these methods are utilized to mimic the speech of a high-ranking government official or a military commander on a large scale, it could present significant consequences for national security."

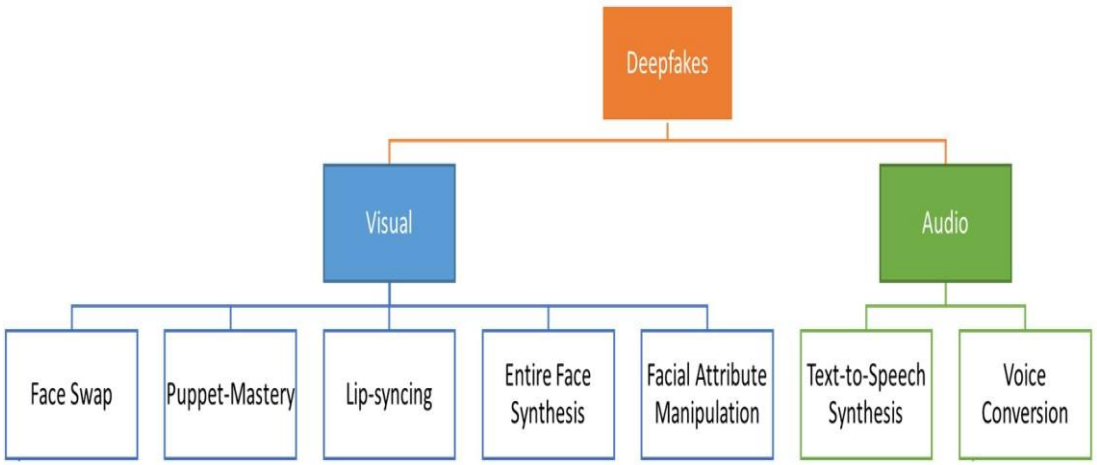

Figure 4 Segregation of Visual & Audio Deepfakes [6]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

The diagram above illustrates that Deepfakes can be divided into two main categories: visual and audio manipulations, depending on the type of manipulation being done. Visual deepfakes can be further classified into different types based on the level of manipulation, such as face swap, lip-syncing, face reenactment, entire face synthesis, and facial attribute manipulation. On the other hand, audio deepfakes are categorized as text-to-speech synthesis and voice conversion.[6]

3.1. Visual Deepfakes – Generation & Detection

3.1.1. Face Swap

Generation

Face swapping with deep learning includes teaching two pairs of encoders and decoders with various faces. These encode and decode the facial characteristics of each face. Following the training, we match the encoder of one face with the decoder of the other. This enables us to combine the characteristics of one face onto the other, producing a new image where one person's face appears as another's. To create realistic swapped faces, we employ methods such as adversarial and perceptual losses. These methods guarantee that the swapped faces look like real faces and display natural expressions. Recent techniques no longer require specific training on individuals, making them more flexible.

Detection

One way to detect face-swap manipulation is by using advanced technology, such as deep learning with CNN and RNN models, to analyze changes in videos. For example, Li and colleagues utilized software to pinpoint facial landmarks and then taught these models to identify fake content in videos. Another method, developed by Guera and team, involves extracting features from each frame and training a unique model to detect deepfakes, particularly in short videos. Li's team exploited the fact that manipulated videos often lack natural eye movements, using a combination of CNN and RNN to spot these distinctive patterns. While these techniques are effective, they may encounter difficulties with videos containing frequent eye blinking or closed eyes. There are various methods, such as those developed by Montserrat and Lima, that employ unique approaches to identify altered content in videos, each having their own advantages and disadvantages.[6]

3.1.2. Puppet Mastery

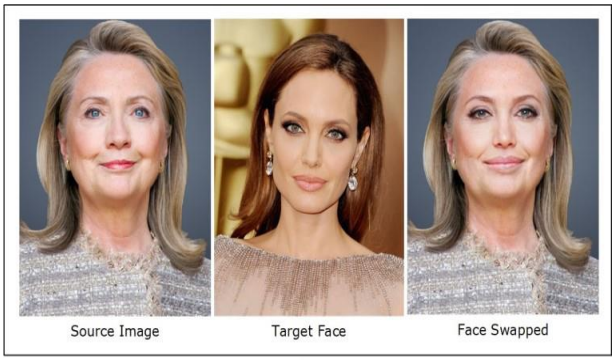

Figure 5 Visual Representation of Faceswap [6]

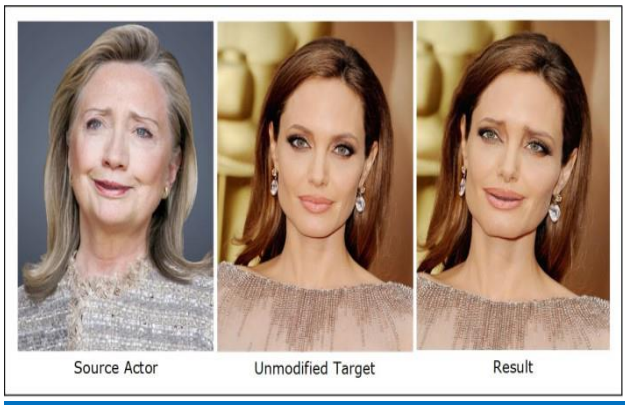

Figure 6 Visual Representation of Puppet Mastery [6]

Generation

Face reenactment, also known as puppet mastery, involves manipulating a person's facial expressions in a video to mimic those of another person. This technique is commonly accomplished using advanced deep learning methods such as CNN and GAN. For instance, Thies and colleagues used real-time 3D facial modeling to transfer facial expressions from one person to another in a video. Face2Face took this a step further by allowing real-time adjustments

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

of facial movements in standard webcam videos, combining 3D modeling with image rendering. GANs have also shown success, with methods like Pix2pixHD producing high-quality images and Kim and team enabling complete reanimation of portrait videos, including changes in head pose and eye gaze. Methods such as ReenactGAN and GANimation utilize boundary latent spaces and emotion action units to convey facial expressions. These approaches enable the production of lifelike reenactment videos by adjusting facial gestures and movements to mirror those of a different individual. This technology has various practical uses in sectors such as film and animation.[6]

Detection

Several deep learning-based techniques have been created to identify deepfake videos created through puppet-mastery in Figure 5&6. Sabir and colleagues utilized a combination of recurrent and convolutional neural networks to analyze the temporal inconsistencies in fake faces, achieving strong results in detecting still frames. Rossler and team integrated both manually crafted and machine-learned features for detection, but noticed a drop in performance when dealing with compressed video files. Afchar et al. utilized simplified CNN models to study the mesoscopic properties of altered content, resulting in a decrease in accuracy. Lastly, Nguyen and team introduced a multi-task CNN approach to simultaneously detect and pinpoint manipulated content, displaying resilience but with a slight decrease in accuracy when faced with new situations. In order to improve performance, Stehouwer and colleagues have developed a Forensic transfer (FT) based CNN method for detecting deepfake videos. This technique is designed to spot deepfakes that involve puppetry by examining the consistency of movements, characteristics, and microscopic details in videos. This method offers a range of options with varying levels of accuracy in detecting deepfakes and computational efficiency. [6]

3.1.3. Lip Syncing

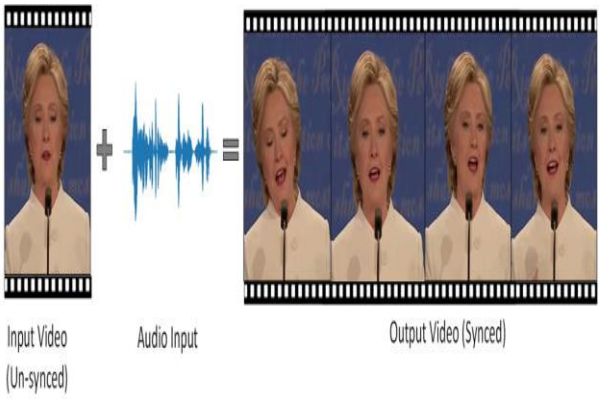

Figure 7 A Visual Representation of Lip-Syncing of an Existing Video to an Arbitrary Audio Clip [6]

Generation

When creating lip-syncing videos, there are different techniques used to synchronize mouth movements with audio. One method involves using recurrent neural networks (RNNs) to match audio features to mouth shapes for each frame, and then adding texture around the mouth using landmarks in Figure 7. Another approach is to use Mel Frequency Cepstral Coefficients (MFCC) features from the audio, which are processed by a CNN-based encoder-decoder to produce a lip-synced video. GAN-based methods use adversarial learning to separate audio-visual representations, resulting in realistic talking face sequences with synchronized lip movement. Some systems capture high-quality 3D facial models of both the source and target actors, reconstructing a 3D mouth model to be applied on the target actor while analyzing the audio channel for alignment. Despite variations, these methods aim to produce convincing lip-synced videos by synchronizing mouth movements with audio input. Advanced technology allows for the creation of lifelike 3D facial models of actors, which can then be used to sync their mouth movements with audio in videos. Despite differences in techniques, the goal remains to produce realistic lip-synced content.[6]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

Detection

Many DL-based methods have been suggested to detect deepfakes. One strategy looks at the discrepancies between phoneme-viseme pairs, noting that certain lip shapes linked to specific phonemes need to be fully closed to articulate them. However, deepfake videos often do not exhibit this characteristic, making it easier to spot them. Another technique involves training a spatio-temporal network with a 3D-CNN ResNet18 feature extractor and a multiscale temporal convolutional network (MS-TCN) on lip-reading datasets, then adjusting the model on deepfake videos. While this method shows promise, it faces challenges with restricted mouth movements in videos. One way to detect deepfake content is by looking for discrepancies between the audio and visual elements. This can be done by calculating a modality dissimilarity score (MDS) to see if they are in sync. A siamese network is used to compare emotional cues in facial expressions and speech, helping to differentiate between genuine and manipulated videos. In addition, a combination of CNN and bidirectional LSTM network is utilized to extract facial features and identify temporal inconsistencies in the content. These advanced techniques are designed to accurately spot deepfake videos by scrutinizing different aspects of both the audio and visual components.[6]

3.1.4. Face Synthesis

Figure 8 Increasingly Improving Improvements in the Quality of Synthetic Faces, as Generated by Variations on Gans.[6]

Generation

Face synthesis, made possible by advancements in deep generative models such as GAN and VAE, involves creating realistic images of human faces, whether they are real or not. These models learn to generate lifelike faces by training on large datasets of facial images. Early models like DCGAN initially improved the quality of generated images, followed by advancements like CoGAN and ProGAN, which further enhanced both the quality and resolution in Figure 8. StyleGAN introduced a mapping network to control various visual features, resulting in cutting-edge high-resolution images with intricate details. StyleGAN2 and TP-GAN continued to improve image quality by addressing issues like artifacts and maintaining identity even with changes in pose. Advancements in AI technology, such as self-attention modules and models like BigGAN and StackGAN, have greatly improved the realism and diversity of generated images. This has led to a significant advancement in the ability to create images that closely resemble real photographs. While these techniques have proven valuable in various industries including gaming, entertainment, and art, there are growing concerns about the potential for misuse in creating fake content and spreading misinformation.[6]

Detection

Advancements in AI technology have greatly improved techniques for detecting image manipulation. Guarnera et al. introduced a method that focuses on analyzing static images using Expectation-Maximization (EM) to train a classifier. While effective, this approach is limited to static images. Nataraj et al. proposed a different method that utilizes pixel co-occurrence matrices and CNNs to accurately distinguish between manipulated and authentic images. Yu et al. created an attribution network architecture that connects input samples to fingerprint images, using correlation indexes for classification. However, this method may face challenges when dealing with post-processing effects such as noise or compression. In a study conducted by Marra

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

and colleagues, they focused on detecting fake images generated by GANs using a multi-task incremental learning approach. Their method is able to adapt to new fake samples generated by GANs, providing robustness in detection. However, it is important to have knowledge of the specific method used to generate the fake content in order to achieve optimal performance. These techniques are helping to advance the field of image manipulation detection by addressing the challenges posed by constantly improving manipulation techniques.[6]

3.1.5. Facial Attribute Manipulation

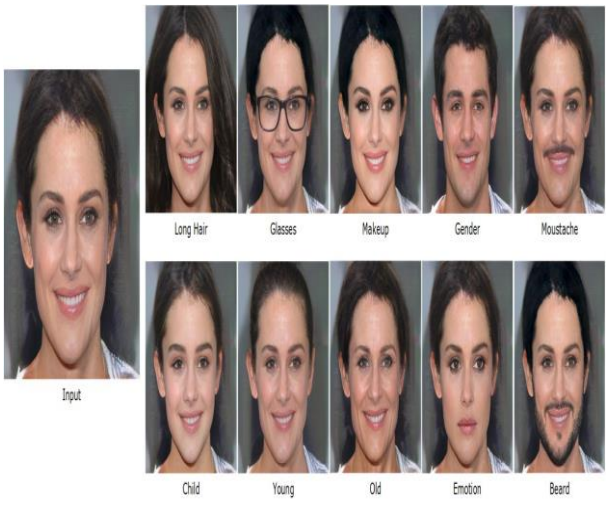

Figure 9 Examples of Different Face Manipulations: Original Sample (Input) and Manipulated Samples [6]

Generation

Manipulating face attributes involves modifying certain characteristics of a face without affecting others. Multiple methods, including those based on Generative Adversarial Networks (GANs), have been created for this purpose in Figure 9. These techniques typically start with an original image of a face and then produce a modified version with changed features like skin tone, hair style, age, or gender. One example is the Invertible Conditional GAN (IcGAN), which utilizes an encoder and conditional GANs to adjust attributes while maintaining the face's original identity, although it may result in some loss of information. Fader Networks, on the other hand, separate image details and attributes directly in a latent space but may lead to some distortion. StarGAN and its upgraded version, StarGAN-v2, allow for switching between different characteristics with a single generator, but could result in some noticeable imperfections. AttGAN integrates attribute classification limitations to guarantee accurate attribute modifications while still maintaining facial features. STGAN targets specific areas related to attributes when adjusting encoded features, whereas approaches like SAGAN and PA-GAN introduce attention mechanisms to pinpoint modifications in designated areas. Together, these methods provide effective ways to alter facial features, although success in maintaining image quality and attribute accuracy may vary.[6]

Detection

Artificial intelligence techniques such as FakeSpotter use deep learning to analyze features from facial recognition models to identify changes in facial attributes, distinguishing between authentic and manipulated faces with support vector machine (SVM) classifiers. However, these techniques may struggle when faced with significant lighting variations. Other approaches either focus on analyzing entire faces or smaller patches, utilizing restricted Boltzmann machines (RBMs) or convolutional neural network (CNN)-based keypoint extractors in combination with SVM classifiers. While these methods demonstrate strong performance, they are vulnerable to attacks that occur after the initial processing. Some methods incorporate attention mechanisms to improve the performance of CNN frameworks, enabling them to detect alterations made by applications like FaceApp or StarGAN. Although these techniques achieve high accuracy by leveraging generative adversarial network (GAN) fingerprint information, they encounter difficulties in removing fingerprints while maintaining the realism of the manipulated faces.[7]

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

3.2. DeepFake Detection Techniques

3.2.1. Deepfake Detection Using MTCNN & RESNET50

Detecting deepfakes is a key aspect of digital media forensics, as it helps identify videos that have been altered to look real. One effective method is using the MTCNN and RESNET50 models. The MTCNN model is great at finding and aligning faces in pictures, ensuring that the analysis focuses on the correct parts. On the other hand, RESNET50 uses its deep residual learning framework to analyze these faces and recognize common patterns found in deepfakes. This duo provides a strong set of tools for spotting manipulated videos by carefully examining facial features. The method is easy to use and highly efficient, serving as a valuable tool for individuals seeking to comprehend and participate in combating online deception. In this Research Paper, we will be implementing this algorithm only as we move ahead to ascertain the accuracy of this technique.

3.2.2. Deepfake Detection Using 2-Phase Learning Architecture

Omkar Salpekar is studying how to identify DeepFakes, which are fake images created using AI to spread false information on social media. The research involves building a binary classifier in two steps. Initially, a specific type of neural network known as a Siamese Network is used to compare the characteristics of real and fake images. This network is trained to minimize discrepancies among features of the same category (real or fake) and to accentuate disparities between different categories. Next, a simpler Convolutional Neural Network (CNN) is integrated to determine whether an image is authentic or fabricated based on these characteristics. The model was trained on a dataset that purposely had an imbalance to mirror the abundance of fake images in real-world situations. By making precise tweaks and refining the architecture, including adjusting learning rates and batch sizes, the model effectively distinguishes between real and fake images with great accuracy. The study also proposes potential enhancements for the future, such as integrating audio indicators and implementing Recurrent Neural Networks (RNNs) for detecting fake videos, in order to improve the model's functionality. In conclusion, this research offers a hopeful solution for uncovering DeepFakes, ultimately crucial for upholding trust and precision in online content.ss4]

3.2.3. Deepfake Detection Using MESONET

MESONET, a cutting-edge deep learning framework, leads the way in detecting DeepFake videos, a prevalent type of digital manipulation where AI substitutes a person's face with someone else's likeness. Created by Darius Afchar et al. in 2018, MESONET has become a critical tool in digital forensics, tackling the growing issue of identifying complex visual forgeries. Unlike traditional methods, MESONET analyzes small-scale features, focusing on subtle details that larger approaches may miss. This refined focus allows MESONET to identify anomalies present in DeepFakes, such as inconsistencies in facial features and texture patterns, which are often disregarded by other methods. MESONET utilizes a powerful Convolutional Neural Network (CNN) architecture to process video data efficiently, ensuring accurate analysis. By focusing on mesoscopic features, MESONET can differentiate between real and manipulated content using a diverse dataset of authentic and forged images. This extensive training enables MESONET to detect minor discrepancies that may indicate DeepFake manipulation, even under challenging conditions like heavy compression. Additionally, MESONET features a flexible design that allows for easy integration into various types of applications and workflows. Its versatility enables both researchers and practitioners to utilize MESONET's abilities in their specific fields, whether it be for academic research or practical digital forensic analyses. Moreover, MESONET provides an Application Programming Interface (API) that simplifies

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

access to its detection features, giving developers the ability to seamlessly integrate MESONET into their software solutions. When it comes to performance, MESONET has shown remarkable accuracy rates exceeding 88% in various studies, even when faced with challenging conditions like significant data compression. This impressive precision, along with MESONET's quick processing speed, highlights its effectiveness as a key defense against the spread of DeepFake content. In a time where digital misinformation threatens societal discussions and honesty, MESONET stands out as a ray of hope, offering a dependable and effective way to detect and fight deceptive visual content. Its impact on digital forensics and beyond is significant, marking the beginning of a new era of trust and transparency.[5]

3.2.4. Deepfake Detection Using Audio-Visual Inconsistency

The process of identifying DeepFakes by analyzing inconsistencies between audio and visual elements in a video is a complex method. DeepFakes are created by smoothly blending one person's face onto another person's body, but they often fail to perfectly match the altered visuals with the original audio. This approach takes advantage of this weakness by using different methods to examine both the visual depiction of speech movements and the corresponding sounds. A method that can be used involves breaking down the video into phonemes, which are the building blocks of speech sounds, to identify discrepancies in mouth movements that do not align with the spoken words. Moreover, the audio-visual coupling model (AVCM) evaluates how closely mouth frames match up with speech segments, where lower similarity scores suggest the possibility of DeepFake manipulation. Contrastive learning techniques also improve detection by comparing the representations of audio and visual segments for verifying identity. An important benefit of this method is its potential for better accuracy compared to approaches that rely solely on visual cues. Furthermore, adding audio data makes it easier to understand and interpret detection evidence subjectively, leading to better comprehension. The efficiency is also enhanced with phoneme-based segmentation techniques showing superior performance in processing video content. However, there are still challenges to overcome, like the need for high-quality datasets for optimal results and the increasing complexity of DeepFake technologies that challenge detection capabilities. Despite these challenges, the continuous advancements in audio-visual inconsistency detection highlight its potential as a powerful tool in fighting against deceptive digital media. As research in this field advances, improvements in detection techniques have the potential to enhance defenses against the spread of fake content.[5]

4. Literature Review

The DeepFakes has caused a havoc in the Content Verification Industry as no one has been spared with its impact and many of us are under an unforseen threat. Former U.S. president Donald Trump posing with Black voters, President Joe Biden discouraging people from voting via telephone or the Pope in a puffy white jacket: Deepfakes of videos, photos and audio recordings have become widespread on various internet platforms. With the right prompt fine-tuning, everyone can create seemingly real images or make the voices of prominent political or economic figures and entertainers say anything they want. While creating a deepfake is not a criminal offense on its own, many governments are nevertheless moving towards stronger regulation when using artificial intelligence to prevent harm to the parties involved. According to the verdict of Mr. **Pavel Goldman-Kalaydin (Head of AI/ML at Sumsub) "**The rise of artificial intelligence is reshaping how fraud is perpetrated and prevented. AI serves as a powerful tool both for anti-fraud solution providers and those committing identity fraud. Our internal statistics show an alarming tenfold increase in the number of AI-generated deepfakes across industries from 2022 to 2023. Deepfakes

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

pave the way for identity theft, scams, and misinformation campaigns on an unprecedented scale**”.** 4,500% increase in the number of frauds attempts in the Philippines over a period of one year is one of the significant changes that have been recorded. It was followed by countries such as Vietnam, United States and Belgium. There is a possibility that deepfake fraud attempts might also start being made in other arenas due to advances being made in artificial intelligence technology such as AI Video Generator Sora. In the following literarture Review, we will disucussing various deepfake detection techniques which are already proposed.These Detection techniques , however are somehow became obsolete because of the current advancements in the technologies used for generating deepfakes. Extensive progress in generating DeepFakes, -notably through Generative Adversarial Networks (GANs), Variational Auto-encoders (VAEs) and recent diffusion models has been witnessed in this field. Moreover, such technologies have made synthetic content more real making detection very difficult using the old approaches. Correspondingly, most prior techniques in deepfake discovery are irrelevant. The detection approaches were using mere sight indications such as blinking awkwardly or unusual hand characteristics but are unsuitable when it comes to latest generation artificially created voices on social platforms. There exist challenges in keeping information from being falsified. This is due to deepfake technologies advancing faster than detection mechanisms which should guard against them. In order to address this, new detection approaches are being explored by researchers, some of which are built on transformer models and biological signals that could provide better resilience against the most recent DeepFake generation techniques. Nevertheless, across various datasets, their accuracy may decrease implying necessity of additional studies aimed at designing simpler algorithms. Some of the previously proposed DeepFake Detection Algorthims are discussed in the subsequent Sections.[6]

4.1. 2-Phase Learning Architecture

4.1.1. The Architecture of 2 Phase Learning for Detecting Deepfake is Sophisticated as it Has 2 Steps

First Phase: The phase uses a ResNet based CNN (Convolutional Neural Network) which is trained like a Siamese Neural Network1. Here it is supposed to scrutinize the input pictures and pull-out distinctive characteristics that will enable differentiating original images from forged ones. Siamese Neural Networks are particularly effective for this task because they are designed to compare pairs of inputs and learn from the differences or similarities between them.

Second Phase: During this part, a 2-layer CNN takes control. It uses the feature encodings got from the first part and analyses them to make a final decision on if an image is authentic. A binary classification is output showing if the image is REAL or FAKE. The proposed architecture is an extension of earlier studies and has proven to be very accurate during testing, successfully recognizing complex GAN-created DeepFake pictures. In this study binary image classification is emphasized in which an image serves as the input for predicting its authenticity Figure 10.

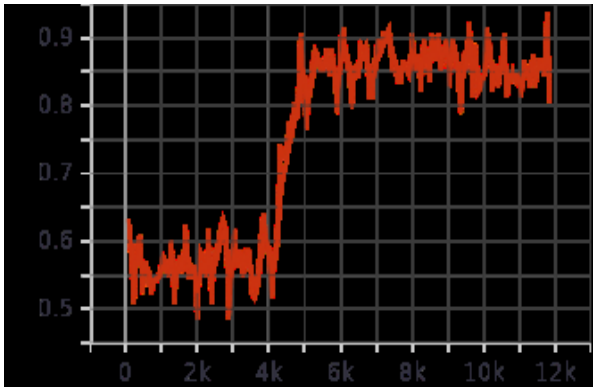

Figure 10 Training Accuracy [6]

The Model has a final training accuracy of 94%. What we have here is just an illustration that the model has undergone effective training hence with high accuracy, it can recognize real images as well as false images. Additionally, we reported

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

the loss values during training as well as validation: training = 0.14, validation = 0.17. When the loss is low then it means that this is not an overfitted model but one that can generalize well hence important consideration for use in other fields where images may be found outside those used during learning sessions Figure 11.

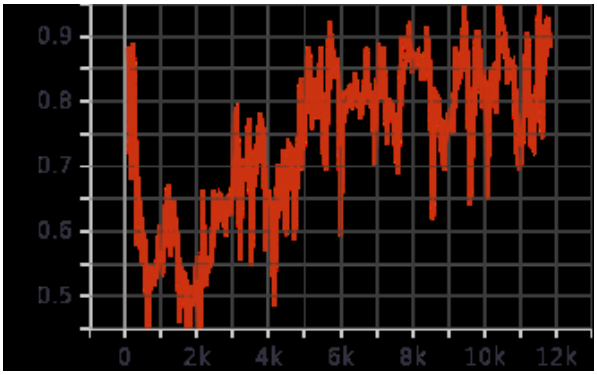

Figure 11 Validation Accuracy [6]

The Validation Accuracy came around 91% which clearly shows the rate of acceptance of data. Apart from this the DFDC DataSet(DeepFake Detection Dataset Challenge) is used for testing DeepFake Images which is around 470 GB of DataSet containing set of Images and Videos(Mp4 Format). Validation accuracy is an important factor as it indicates how well a model may perform on new and unfamiliar input samples – a point of great concern in practical utilization. By so doing, it is plausible that a model will avoid memorizing the train data only but would rather assimilate deeper fakes attributes for better performance under real environments. These numbers emphasize how fast the subject of rising to understanding DeepFake is evolving as well as how much capability deep learning models have when it comes to averting propagating untrue information. Even so, there is a need for continual refining of these kinds of models and checking them against the latest DeepFake methods so that they remain correct and dependable.

Second Phase: To detect DeepFakes a full-on approach is used in phase 2 of training. Initially, the Convolutional Feature Fusion Network (CFFN) is trained with a Siamese Network to enhance its parameters. In the network there are powerful features including but not limited to occlusions, pixelation around the face, color gradients and abnormal shadowing which distinguish this type of fakes. Real images have the potential to create many training cases by combining various fake images in a dataset predominantly biased towards imaginary illustrations. After which, a small Convolutional Neural Network simply known as the Classifier Network is affixed to the CFFN. These concatenated encodings are passed input through this classier network from the last convolutional levels in CFFN. It includes more linear and convolutional layers that produce binary class decisions. Both the CFFN and the Classifier Network together make the whole structure that is trained with cross-entropy loss. Through this approach,two classes are effectively separated from each other due to their dissimilarities in characteristics. While the Classifier Network phase focuses on fine-tuning the earlier layers of the CFFN and training the final two optimal performance layers, Siamese Network phase directs its efforts to establishing sturdy feature detectors. What caters for the working of this model is TensorFlow.[4]

4.2. Mesonet: A Compact Facial Video Forgery Detection Network

Mesonet, an AI system, is designed to identify counterfeit facial videos, most notably those created through Deepfake and Face to Face technologies. It comes from Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen.

4.2.1. Key Features

Compact Architecture: MESONET is renowned for its compactness, with only a few layers. The design is intended to hasten detection of minute anomalies evident in falsified videos because of its ability to examine mesoscopic information in images.

High Efficiency: It has been optimized for high-

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

speed video processing that ensures quality results.

High Detection Rate: MESONET has demonstrated a very successful detection rate, with more than 98% accuracy for Deepfake and 95% for Face-to-Face forgeries.

4.2.2. Implementation

Data Preprocessing: The initial videos are preprocessed before being input to the MESONET model. The very first step is extracting frames from the videos and then resizing them to a uniform size, typically 256x256 pixels. This step is important since it ensures that the network gets inputs in the same format hence helping to preserve accuracy of detection.

Model Architecture: MESONET is based on a shallow convolutional neural network (CNN) contrary to other models of deep learning that are much more in-depth. The developers chose this way of design for the network specifically for it to concentrate on mesoscopic attributes of the images such as textures and patterns that are quite too small for a human to see, but can be identified by any CNN. Four convolutions along with batch normalization with ReLU after every fully connected layer are forming an artificial neural network while two dense ones follow them immediately. The last layer employs a sigmoid activation function to depict whether the input is likely to be a Deepfake or not by giving us his probability score.

Model training: we use a labeled dataset in a training process where processed frames are fed as input and marked showing the reality or fakeness of each. This model is energized by a binary cross-entropy loss function meant for binary classification tasks. Throughout the training phase, the model is taught how to detect slight discrepancies and other signs of deepfake footages.

Detection and Evaluation: MESONET can be used to evaluate new video frames after it has been trained on them. For every frame, the model gives a probability score that helps classify it as real or fake based on a predetermined threshold. Commonly observed measures in evaluating the performance of MESONET are accuracy, precision, recall and F1-score. This provides insights into its overall effectiveness in detecting Deepfakes.

Deployment: For real application, MESONET can be used by connecting it with other security systems, which help to study movies of Deepfakes. The model can work with video streams live or analyze saved movie files. Its design is small-sized and effective in terms of processing power, which makes it easy to incorporate it in many different providences such as news sites including social networks and content moderation systems.[5]

4.3. Audio Video Inconsistency

The Audio-Video Inconsistency Deepfake Detection Algorithm is a new way to spot fake videos by finding differences between what you hear and what you see on screen with a particular emphasis placed on finding out how well sounds sync with lips. This technique is very important in detecting those deepfakes which preserve faces but edit lip motion to suit sound.

4.3.1. Key Features

Temporal Inconsistency Analysis: The routine goes through the audio and visual lip movement data of your video to understand how they align in time while isolating small inconsistencies human beings may not notice.

Biological Link Exploitation: This program is created so that it can think like people do which these days could mean appropriating those areas where lips move together with head movements thereby provoking better detection rates.

4.3.2. Performance

High Precision: It has demonstrated an average accuracy of 95.3%, significantly surpassing baseline techniques in identifying lip-syncing deepfakes.

Real-world Application: Achieves real-world accuracy of up to 90.2% for example in video calls showing robustness and practical deployment capability.

4.3.3. Implementation

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

Gathering and preprocessing data: The initial phase entails gathering a vast array of video samples having authentic and counterfeit aspects. It involves collecting video samples with both real and fake substance in them. Next, preprocess these videos by aligning their audios with videos so that they can be analyzed together. The next stage after that could be taking out lip area from frames before translating the sound into something else that would be used while extracting features.

Feature Extraction: After preprocessing the data, the algorithm looks for important elements from both sound and movie parts. When it comes to the visual one, it centers on how lips move; while in the case of listening element, the algorithm analyses phonetic sounds related to lip movements. The aim is to get features that would remain the same for both types of videos.

Temporal Analysis: Time matters, so the features obtained are analyzed temporally to verify if there are discrepancies in the sound or video tracks. This process entails timing of lip motion against the respective sound in the recording. The algorithm seeks out discontinuities such as variations in speed, sudden changes in movement patterns or any other hints that there might be some fraudulent activities between sound and image sequences.

Machine Learning Model Training: The process of training a machine learning model involves using the extracted features, typically a neural network is used. The training process includes feeding labelled data to the model, real examples being positive and fake samples negative. When the model is trained using such data sets it can separate regular from irregular data points.

Detection and Labeling: After learning from data, our in-house experts use this methodical tool for admitting whether it is coming from genuine or counterfeit sources by looking at its pictorial coherence with other audios. Application for checking this is done in another means: using the same geometrical features derivation with subsequent stress as a tunnel; it is made plausible by these teachings to forecast whether a video is fake or not.

Evaluation and Optimization: In the final step, evaluate the algorithm using precision, recall, as well as F1 score amongst other metrics to understand its performance. These evaluation outcomes will guide whether further optimization is needed through modifying the model parameters or by changing how we extract features.[5]

4.4. RESNET 50 and MTCNN Deepfake Detection Algorithm

4.4.1. RESNET 50

Introduction: The ResNet model has a variant called ResNet-50 that helps in image recognition tasks. It was first introduced by Kaiming He and team in their 2015 paper titled “Deep Residual Learning for Image Recognition”. This name ‘50’ was named due to the total layers known as so that it could learn large quantities of information without facing issues associated with disappearing gradients.

Architecture: The architecture of ResNet-50 consists of 48 convolutional layers. Additionally, there is a MaxPool layer and an average pool layer in the architecture. Shortcut connections are the key concept in ResNet-50. These connections allow for skipping some layers when they are not significant for calculations. Thus, deep networks are successfully trained since this method addresses the problem of disappearing gradients. It is also worth mentioning that the arrangement of these layers is bottleneck which implies lower number of parameters and hence better efficiency of training process.

Implementation: ResNet-50 is commonly implemented in deep learning frameworks such as TensorFlow and Keras. The procedure includes starting the network with pre-trained weights and fine-tuning on a particular dataset. This new task transfer strategy makes use of existing models that have seen large features in an open domain example like ImageNet as the basis for generalizing them after encountering another set

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

through another way adopted by nets.

4.4.2. MTCNN (Multi-Task Cascaded Convolutional Networks)

Introduction: MTCNN is an advanced algorithm for face recognition, which is very useful for face detection as well as for face alignment tasks in images and other types of artworks. It's mostly applied to the systems where high level of accuracy in identifying individuals' faces like detecting deepfake video clips is required.

Analysis: The MTCNN detects faces by scrutinizing an image at various resolutions via scanning. Through a three-network layer that makes predictions based on face location as well as other landmarks through coarse-fine iterative means, this detection network tracks faces excellently under different conditions including those with poses.

Finding deepfakes: MTCNN could potentially be utilized for the extraction of faces from video frames when we talk about deepfake detection. The verification whether such faces are authentic or been doctored is carried out through analyzing with deep learning algorithms. It's this specific method's efficiency at picking up on minor irregularities within facial aspects that renders it such an invaluable asset when it comes to recognition of fake media sources.

5. Methodology

When you start our AI way of discover deepfake, firstly gather a number of different real-looking ones in contrast with computer generated fakes by using pictures which are not situated in real life but seem realistic at first glance before becoming suspicious since everything around may turn out being part of simulation as well as being an imitation without informing anything definite; this is why they do not have any meaning except for being equal among themselves visually although you can't understand why exactly this picture was made here or there in order to be precise while having several representations of such a kind - none makes more sense than other ones being in ideal agreement together at some point however if we consider the different periods or events it refers then we'll find out that each of them has own relation with some kind of reality although would be impossible otherwise even though implies merely existential nature. After this, an adept pattern recognition tool known as ResNet-50 v1 is used to look at facial images very carefully in order to recognize the authentic ones by looking at some details that only show their truthfulness. Consequently, this information is used to teach a computing system how to tell the difference between what is fake and real photograph. To show that it works well enough, we use some statistical techniques for testing the ability of our system that can differentiate actual damages caused by Deepfakes from images that exist in reality. Our aim is to develop a reliable system of authentication for pictures that are false for the purpose of securing pictures on the internet against unauthorised distortions.

5.1. Data Collection

In order to implement the algorithm and tests it's accuracy, we need a large amount of data.Therefore We considered the Dataset of "OpenForensics: Multi-Face Forgery Detection And Segmentation In-The-Wild Dataset [V.1.0.0] " as our testing purpose .The dataset contains around 5000 images divided into 2 categories, one is of real images and another one is of fake images.The size of the dataset is about 2.6 GB which is efficient for proving the efficiency of algorithm in Figure 12.

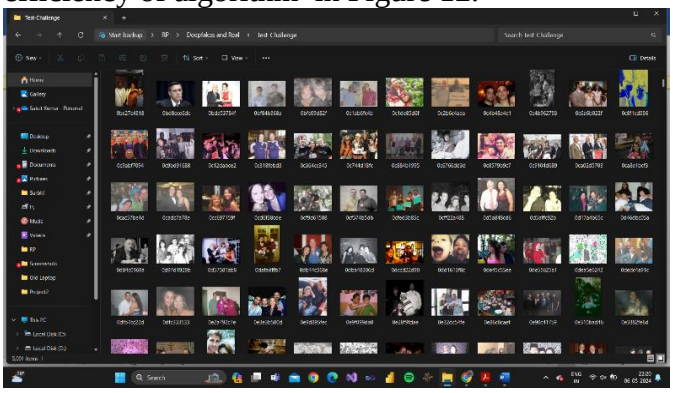

Figure 12 The Dataset [10]

It is to be noted here that we are here implementing an already proposed algorithm and testing its accuracy and not proposing any new

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

solution because we can only get new solutions have improving in previous ones in Figure 13.

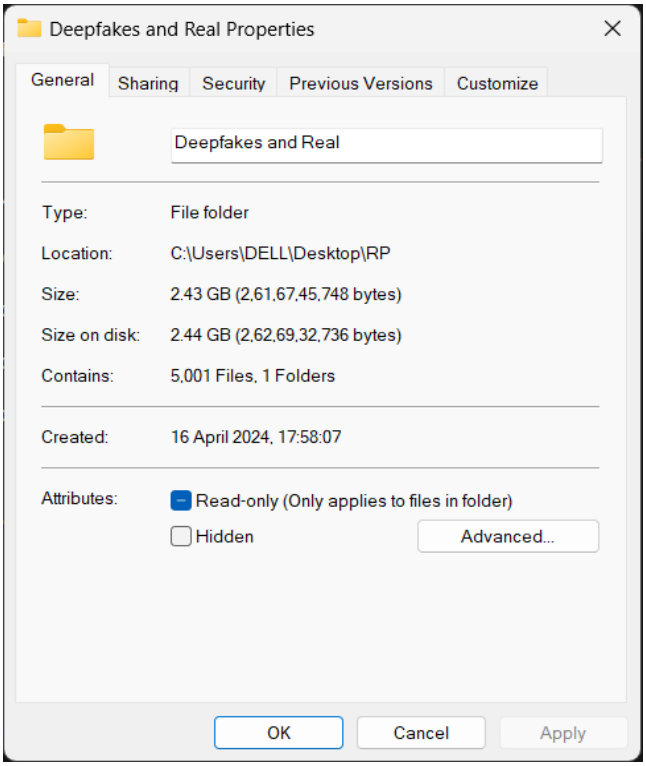

Figure 13 Properties of Dataset

5.2. Preprocessing

Initialization: Initialize the MTCNN detector and the ResNet-50 model. With the loading of pre-trained ImageNet weights in the ResNet-50 model and configuring it to exclude the top layer.

5.2.1. Pre-Processing for RESNET 50

- **Image Loading:** Each image will be loaded with 224x224 pixels as the target size expected by ResNet-50.
- **Image Conversion:** After the image is loaded, it shall be converted to an array format.
- **Dimension Expansion:** For the model to process the image, the image array shall have an extra dimension to it which signifies the number of images.
- **Preprocessing:** Before the image array is fed into the ResNet-50 model, we will process it by taking advantage of Keras preprocess input function, which will normalize the pixel values in a sense supposed for the input of this model.

Feature Extraction: Once we have processed the images when the feature extraction is ready the ResNet-50 model as a feature extractor devoid of its top classification layer is used. This model reads in input images then produces for each image a vector of features that are necessary characteristics in order to detect fake videos later.

Output Preparation: Preparing the final result the output vectors of features obtained from "ResNet-50" are next subjected to flattening in readiness for the next stage of our pipeline detection of deepfake. Achieving this involves putting features in a structure that classifies it as suitable enough to feed to our classifiers, which shall be elaborated on further in this method.

5.3. Feature Extraction Using RESNET 50

RESNET 50 is a 50 Layer Deep Convolutional Neural Network designed for image recognition and classification. It utilizes residual learning to ease the training of very deep networks.

5.3.1. Feature Extraction Process

- **Input Image:** At first, the process necessitates an image input for classification or examination.
- **Convolutional Layers:** An image goes through various convolutional layers aimed at extracting characteristics. Each layer takes care of different parts of the image.
- **Residual Blocks:** The ResNet-50 has its core in something called Residual Blocks, each of them having a shortcut which just goes over one or several layers.
- **Bottleneck Design:** It has a bottleneck design which employs convolutions of 1x1, 3x3 and 1x1 in its residual blocks to reduce the number of parameters.

5.3.2. Extracting Features

- **Intermediate Layers:** Intermediate Layers are responsible for extracting features from

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

image data when they pass through the network.

- **Global Average Pooling:** In the end, a global average pooling layer is used to reduce spatial dimensions without losing vital information.
- **Output Features:** The extracted features from an image are then represented by the output feature vector at last layer.

5.4. Face Detection Using MTCNN

MTCNN stands for Multi-Task Cascaded Convolutional Networks.It's a deep learning model designed for face detection, which also provides facial landmarks.

5.4.1. How MTCNN Works?

Stage 1: Proposal Network (P-Net)

- A shallow cnn examines the image to propose areas on the face.
- We will shrink the picture to various sizes to cater for face sizes in a range of scales (a picture pyramid).
- Shows the probable bounds boxes and scores which determine their confidence.

Stage 2: Refine Network (R-Net)

- It enhances the boxes proposed by P-Net.
- It gets rid of many false alarms.
- It does further convolution to improve the accuracy of detection.

Stage 3: Output Network (O-Net)

- Outputs facial landmarks using a complex CNN to ensure the utmost precision.
- Provides the last bounding box refinements.
- Outputs facial landmarks, for example, eyes, nose and the mouth.

5.5. Classification Algorithm

In the following section, the algorithm that was implemented in order to classify between DeepFakes and Real Images is given. The following algorithm clearly classifies between the two constraints and shows an accurate result.

- **Algorithm:** Detect and Classify Real or Fake Images
- **Input:** Lists of paths to real and fake images
- **Output:** Performance metrics (Accuracy, Precision, Recall, F1 Score), Hypothesis Testing Result

BEGIN

1. Initialize the MTCNN detector and the ResNet50 model.
2. Define a function to preprocess images for ResNet50.
3. Define a function to extract features using ResNet50.
4. Define a function to detect faces using MTCNN.
5. Load the dataset of real and fake images.
6. Define a placeholder function to classify images as real or fake.
7. Define a function to analyze results and calculate performance metrics.
8. Define a function for hypothesis testing between real and fake groups.
9. For each image path in the real images list:

a. Preprocess the image.
b. Extract features.
c. Classify the image and store the result.

10. Repeat step 9 for fake images list.
11. Analyze the results to calculate Accuracy, Precision, Recall, and F1 Score.
12. Perform hypothesis testing with the classification results of real and fake images.
13. Print the performance metrics and hypothesis testing result.

End

5.6. Performance Metrics

Accuracy: This metrics calculates the proportion of true results among the total number of cases examined. It is the ratio of correctly predicted observations to the total observations.

The Formula is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Precision: This metric refers to the ratio of true positive observations to the total predicted positives. It measures the quality of correct predictions.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

The Formula is:

$$\text{Precision} = \frac{\text{True Positives}}{\text{Total Number of Predictions}}$$

Recall: This metric calculates the ratio of true positive observations to all actual positives.

The Formula is:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

6. Implementation

This section of this research paper defines description of the system setup uisng a suitable diagram, discusses the coding and implementation process of the deepfake detection algorithm and sheds light on the challenges faced during the implementation in Figure 14.

6.1. System Architecture

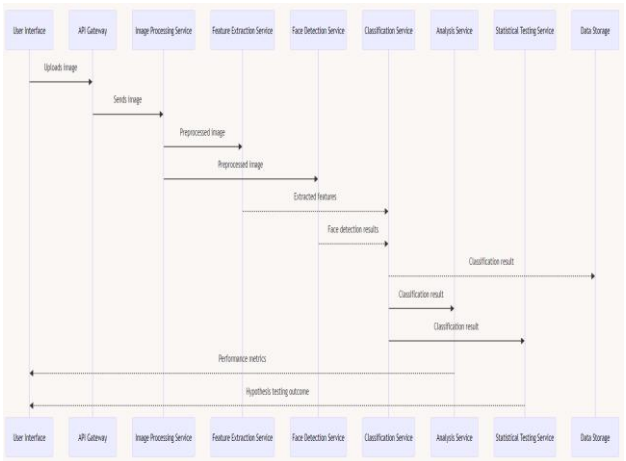

Figure 14 Architecture

Flow of Operation in the above image:

- The image is uploaded by the user using the UI. Via the API Gateway, the image goes to the Image Processing Service.
- When the image is eventually delivered to the service, first it is preprocessed before being sent to both the Feature Extraction Service and Face Detection Service.
- Finally the Classification Service gets both face detection results and the extracted features.
- It also stores the classification result which is then sent for analysis at Analysis Service and Statistical Testing Service.
- It's the Analysis service that will determine the performance metrics.
- The Statistical Testing Service runs the hypothesis test and provides the user with feedback using the UI, including performance metrics and hypothesis testing outcome.

6.2. Algorithm Implementation

Installing a system to detect and classify real and fake images including several steps that combine different machine learning methods and statistical analyses. Initially, two main models must be initialized by the system: the MTCNN file and the ResNet50 that is used to extract features. MTCNN easily locates facial features in images, and ResNet50 lets details to be extracted without upper classification layer from images using already learned from ImageNet database parameters. Once we have initialized our models, we define an image preprocessing function that guarantees they are correctly styled for usage with the ResNet50 model; this includes resizing them down to 224x224 pixels in size plus applying any specific manipulation procedures dictated by its architecture before we proceed with our next step – building a feature extraction procedure which processes these pre-processed images through the ResNet50 classifier so that we get out what each image conveys in terms of features. Meanwhile, 'create face detection function using MTCNN model'; In this regard, this function is crucial because it helps isolate genuine from counterfeit pictures. Having developed both pre-processing along with feature extraction processes, you load together both real and fake datasets from an approved location. Further these images need their paths indicated before those programs can open them respectively. Evaluating classification function performance means developing a result analysis function capable of estimating performance metrics like accuracy, recall, precision, F1 score among others. This will be

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

done by another, more advanced function that would correctly assign appropriate tags to the images in terms of extracted features. Consequently, we define a placeholder function in the class for telling whether an image is authentic or phony. These measurements give a full evaluation of the model's effectiveness. oreover, the paper presents hypothesis testing features to make statistical comparison between actual and false groups. To find out whether there are any great differences between these two collections, a t-test is carried out. Following this, the program goes through every photograph in datasets of genuine and forged pictures and performs preprocessing, feature extraction and classification. From these results, classification performance parameters are calculated. The machine produces results of hypothesis tests and performance metrics, which establish that it can be able to differentiate a true picture from a fake one. Every minute action in creating, detecting, or classifying images is done ranging from manipulation and collection of photos to their validation through statistical expressions.

6.3. Challenges and Solutions

Several challenges and limitations may arise during the implementation of the detection and classification process, along with potential solutions:

Model Initialization and Memory Constraints:

- **Challenge**: Loading large models like MTCNN and ResNet50 requires significant memory, which could be a limitation on systems with limited resources.
- **Solution**: Optimize memory usage by loading models once and reusing them, or consider using a cloud-based platform with scalable resources.

Image Preprocessing:

- **Challenge**: The preprocessing function assumes all images are of a uniform size and format, which may not be the case in real-world scenarios.
- **Solution**: Implement additional preprocessing steps to handle various image sizes and formats, including error handling for corrupt images.

Feature Extraction:

- **Challenge**: The feature extraction process is computationally intensive and could be slow for large datasets.
- **Solution**: Utilize GPU acceleration for faster computation, or batch processing to optimize resource usage.

Face Detection:

- **Challenge**: MTCNN might produce false positives or miss faces, affecting the subsequent classification accuracy.
- **Solution**: Fine-tune the face detector or use additional face detection methods to improve accuracy.

Data Loading:

- **Challenge**: The dataset loading mechanism is a placeholder, and actual implementation may face issues with data access, especially with large datasets.
- **Solution**: Use data generators or streaming mechanisms to load data in chunks, preventing memory overflow.

Classification Logic:

- **Challenge**: The classify_image function is not implemented, and designing an accurate classifier is complex.
- **Solution**: Develop a machine learning model trained on a labeled dataset, or use transfer learning techniques to adapt existing models.

Performance Metrics Calculation:

- **Challenge**: Performance metrics assume binary classification, but in practice, there could be more nuanced categories or multi-label classification.
- **Solution**: Extend the metrics calculation to handle multi-class or multi-label scenarios, and ensure balanced datasets to avoid skewed metrics.

Hypothesis Testing:

- **Challenge**: The hypothesis testing function assumes that the data groups are normally

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

distributed and independent, which may not hold true.

- **Solution**: Verify assumptions before hypothesis testing, and consider non-parametric tests if assumptions are violated.

Scalability and Speed:

- **Challenge**: The loop for detection and classification might be slow for large numbers of images.
- **Solution**: Parallelize the loop using multi-threading or distributed computing techniques to increase throughput.

Error Handling:

- **Challenge**: The code lacks error handling, which could lead to crashes or undefined behavior.
- **Solution**: Implement robust error handling and logging to manage exceptions and ensure system stability.

Model Generalization:

- **Challenge**: The model might not generalize well to unseen data or different domains.
- **Solution**: Use a diverse training dataset, perform cross-validation, and continuously update the model with new data.

Ethical Considerations:

- **Challenge**: The use of facial recognition and classification raises privacy and ethical concerns.
- **Solution**: Implement strict data handling policies, obtain necessary consents, and ensure compliance with relevant regulations.

7. Results & Discussions

The Algorithm implemented mainly two models that are MTCNN for Face Detection and RESNET 50 for Feature Extraction. The Images given as Input where passed through the 50-layer convolutional layers of these models so that relevant informations regarding the image can be formulated and retrieved. During the implementation of the algorithms, we passed the images from the dataset downloaded .The Size of the dataset was 2.43 GB and contained around 5000 images . The images were large in number and since python language has a slow compiler, we distributed the images into batches of 50 at a time. This was done because during the iteration of images, 5000 images at a time won't be possible as during it's further processing each image has to pass through 50 Layer CNN. So, to save time and avoid space complexity, a batch of 50 images at a time is passed. Each image either results in True or False. The Results showed that most of the images which were a bit distorted or had a blurr effect, gave output as Fake. It is because the model recognizes it as no face detected internally and then gives Fake as an Output.

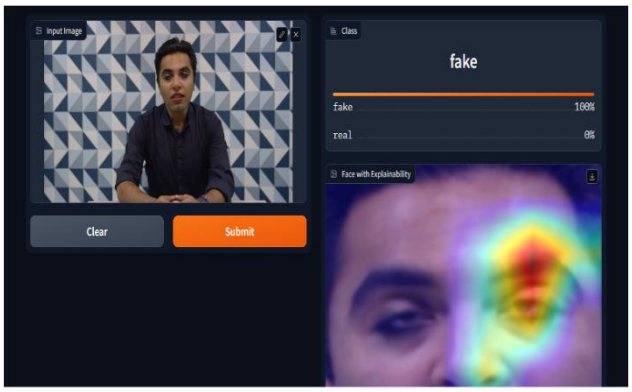

Figure 15 Testing

As we can see in the above Figure 15, the background of the image is distorted. Not only this, the effects on eyes are a bit artistic which also shows that the image has been generated using an AI tool. Therefore, with complete 100% accuracy, the image is fake.

7.1. Consideration of Hypothesis

In regard with experiments carried throughout this research paper, we can consider this Hypothesis.

Assumptions

H_0 – Current DeepFake Detection Techniques are less effective in detecting deepfake images due to their limitations.

H_1 - Current DeepFake Detection Techniques are NOT less effective in detecting deepfake images due to their limitations.

Values

Population Mean Effectiveness (mu): This is the expected effectiveness of the deepfake

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

detection algorithm. Here it is around 70% (69.432) =0.70 after testing a large number of images.

Standard Deviation of 'mu' (Σ): This resmebles the variance or dispersion in population mean effectiveness and it came out to **be 10 % =0.10.**

Sample Size (n): Sample Size is the small sample taken from the complete dataset which was taken as 50.

Sample Mean Effectiveness ({x̄}): This is the Mean Value of the Sample Size that is 50 Observations. The effectiveness showed that out 50 samples, 72%=0.72 were correctly identified as deepfakes.

Significance Level (α): This is the probability of rejecting the Null Hypothesis. It means we are willing to accept 5%=0.05 chance of incorrectly rejecting the Null Hypothesis.

7.1.1. Calculation of Z-Score

$$Z = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})}$$

$$Z = \frac{(0.72 - 0.70)}{(0.10/\sqrt{50})}$$

$$Z = \frac{0.02}{(0.10/7.071)}$$

$$Z = \frac{0.02}{0.01414}$$

$$Z = 1.414$$

Figure 16 Calculating Z Score

Determining the Critical Value

At a significance level of 5%, the value of Z that is critical for a two-tailed test is roughly ±1.96. (The selection of two-tailed test stems from the fact that we are considering the likelihood of having a sample mean different from population mean.)

Decision Rule

- If the absolute value of Z is greater than 1.96, reject H0.
- If the absolute value of Z is less than 1.96, do not reject H0.

Conclusion

Since our calculated Z-score of 1.414 is less than the critical value of 1.96, we do not reject the null hypothesis. This means that, based on our sample, there isn't enough evidence to conclude that current deepfake detection algorithms are more effective than what the Alternate hypothesis states.

Conclusion

In Conclusion, with a concentration on Generative AI and Large Language Models, the document provides an extensive review of how artificial intelligence terrain changes. Artificial Intelligence models that are traditional are outclassed by them thereby becoming robots of the future. One of the dilemmas raised in the manuscript are the authenticities of things like DeepFakes which can imitate anything perfectly hence making it almost impossible to tell whether they are true or not. The study provides an in-depth assessment of different deepfake identification methods, and also introduces a new approach which combines the positives from RESNET-50 and MTCNN models. The objective of this particular technique is to improve the reliability in discerning genuine and counterfeit photographs. The least square method is used for hypothesis testing because current systems have little success in recognizing deepfakes that are highly lifelike. The research paper we're discussing about combines neural networks, deep learning and advanced algorithms. It creates a basis for future progress concerning the use of AI in verifying content. What the study builds on is the idea that innovation in artificial intelligence must be ongoing so long as we face complications from DeepFakes; this will guarantee the integrity of digital media remains preserved and people have trust in it too.

References

[1]. Erik Brynjolfsson, Danielle Li, Lindsey R. Raymond (2023), Generative AI at Work,

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 05 May 2024
Page No: 1489- 1511
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0206

Working Paper 31161, National Bureau of Economic Research

[2]. Baidoo-Anu, D., Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. Journal of AI. 7(1), 52-62.

[3]. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[4]. Omkar Palekar, DeepFake Image Detection, CS230, Standford University

[5]. Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, And Andrew H. Sung's "Deepfake Detection: A Systematic Literature Review", DOI No - 10.1109/Access.2022.3154404

[6]. Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward

[7]. https://www.marketsandmarkets.com/Market-Reports/generative-ai-market-142870584.html

[8]. https://appian.com/blog/acp/process-automation/generative-ai-vs-large-language-models.html

[9]. https://ourworldindata.org/artificial-intelligence#:~:text=Just%2010%20years%20ago%2C%20no,at%20least%20in%20some%20tests.

[10]. https://zenodo.org/records/5528418#.YpdlS2hBzDd