Assessing the Effectiveness of Network Security Tools in Mitigating the Impact of Deepfakes AI on Public Trust in Media

Abstract

The rising threat of deepfake technology challenges public trust in media, necessitating robust countermeasures. This study proposes the Anti-DFK framework, a comprehensive strategy to mitigate the spread of deepfakes on major social platforms such as Instagram, Facebook, YouTube, and Twitter. The framework integrates deep learning-based detection engines, digital watermarking, and advanced network access controls, including URL filtering, domain reputation filtering, content-type filtering, and Geo-IP blocking. Analyzing historical deepfake data, user engagement metrics, and public sentiment from Kaggle Datasets the study employed deep learning models—CNNs, LSTMs, and Transformer-based—to evaluate detection capabilities, achieving a highest controlled environment accuracy of 0.97. Digital watermarking techniques were tested for robustness against various attacks, with the DCT method displaying significant resilience. Network access controls were assessed for their effectiveness in curtailing the spread of deepfakes, with content filtering proving most effective by reducing dissemination by nearly 80%. Findings indicate a critical negative impact of deepfakes on public trust, underscoring the need for the integrated approach offered by the Anti-DFK framework. The study concludes that implementing these sophisticated detection tools, combined with robust digital watermarking and stringent network controls, can significantly enhance the integrity of media content and restore public confidence.

Keywords: deepfake detection, digital watermarking, network access controls, public trust, Anti-DFK framework

1. Introduction

In recent years, the advent of deepfake technology has presented a novel and formidable challenge to public trust in the media. These sophisticated digital manipulations of video and audio content, enabled by advanced artificial intelligence and machine learning techniques have become more accessible and their applications more widespread [1]. This new development comes with the potential for misuse in various spheres —political, social, and personal, and so on, necessitating the need for urgent and effective countermeasures.

Recent high-profile incidents involving deepfakes have drawn significant attention to the potential harms associated with this technology, highlighting the urgency for effective countermeasures to safeguard public trust in the media. For instance, explicit deepfake images of celebrity-Taylor Swift, circulating online which depicted her in highly inappropriate and sexualized scenarios that she had never participated in, were viewed millions of times online, highlighting the profound personal and societal consequences of deepfakes. These images, created without her consent, utilizing advanced AI technologies to manipulate her likeness, which resulted in significant emotional and reputational harm have prompted U.S. legislators to call for new laws to criminalize the creation and distribution of such content, reflecting a growing recognition of the need for robust legal and regulatory frameworks [3].

In addition, the use of deepfake technology in political contexts, such as the generation of misleading robocalls mimicking public figures to suppress voter turnout, illustrates deepfakes' potential to undermine democratic processes. This was notably seen in the 2024 U.S. presidential primaries, where AI-generated voice messages attempted to discourage voter participation by impersonating President Joe Biden. These incidents not only disrupt elections but also contribute to broader issues of misinformation and distrust, complicating the public's ability to discern truth in the media [4]. Sepec [5] avers that there have been increasing instances of deepfakes in cyberbullying and revenge porn, which invade privacy and cause significant emotional distress.

The societal impact of deepfakes extends beyond individual incidents. The pervasive nature of this technology threatens to erode trust in media as a whole, particularly on platforms like Facebook and YouTube, making it difficult for the public to discern truth from fabrication [6]. This erosion of trust can lead to widespread skepticism, impacting public discourse and democratic processes. For instance, a deepfake could potentially alter stock market movements, manipulate election outcomes, or incite violence by spreading false information about sensitive issues [3][4][5].

In response to these challenges, significant development in detection technologies and legal frameworks has been underway. However, these measures alone have proven insufficient as deepfake technology continues to evolve rapidly, outpacing current mitigation techniques. Digital watermarking and network access controls serve as critical layers of defense. Watermarking helps assert the authenticity of digital media, while network controls can prevent the spread of identified deepfakes on platforms like Instagram and X, effectively cutting off the means by which such content can go viral [7]. Despite these advancements, the integration of these technologies on a platform-wide scale remains a complex challenge, with significant technical, ethical, and operational hurdles to overcome [8][9].

Given these developments and the profound implications of deepfakes, there is a clear and present need to assess and enhance the effectiveness of network security tools in preserving public trust in media on these platforms. This study develops and proposes the Anti-DFK framework, a comprehensive multi-layered defense strategy that integrates deep learning-based detection, digital watermarking, and network access controls to mitigate the spread of deepfake content on public media platforms like Instagram, Facebook, YouTube, and X, thereby preserving public trust in these media platforms.

The study achieves four main goals:

1. Evaluation of the capabilities, trends, and societal impact of deepfakes across various media platforms.
2. Investigation of the technical capabilities and limitations of current deep learning-based detection engines for identifying deepfake content.
3. Evaluation of the effectiveness of digital watermarking technologies in maintaining the integrity of original media content on public platforms.
4. Design of network access controls that effectively restrict the dissemination of identified deep fake content

2. Literature Review Structure

Deepfake technology, oftentimes scrutinized and condemned for its misuse around the world, has legitimate applications across various fields that demonstrate its creative and educational potential. According to Kalmykov [11], in the film and entertainment industries, deepfakes are utilized to enhance storytelling, allowing filmmakers to create seamless dubbing in multiple languages or resurrect the performances of past actors. In the realm of art, artists engage with deepfake technology to challenge perceptions of reality and explore new expressions of creativity. Additionally, the educational sectors leverage on deepfakes for more engaging historical recreations or simulations, providing students with immersive learning experiences that were previously unattainable. These distinct applications reveal deepfakes' potential to enrich how we experience media, understand history, and appreciate art, indicating a promising horizon for this technology when guided by ethical use [10][12].

Although deepfakes powerful technology has the potential to revolutionize the media industry as we know, Pawelec [13] state that the abuse of deepfakes in the political arena has emerged as a significant threat to democratic processes, with incidents worldwide illustrating how these technologies are weaponized to manipulate public opinion and disrupt elections. Deepfake technology, particularly through AI-generated audio and

video, has been used to create convincing misinformations regarding high profile political figures, thereby influencing voter behavior and public sentiment [14][15]. For instance, during the recent elections in Slovakia and Nigeria, deepfake audios were circulated to falsely implicate politicians in controversial statements or actions, this was solely done to significantly impact their public image and electoral prospects [16].

Deepfakes technology has proliferated social media platforms and has made it a playground for the circulation of misinformation; this realistic looking manipulated media has significantly implicated public discussion in political, social, and personal settings, as it often blurs the lines between reality and fabrication in ways that can mislead viewers, cause harm and distort public perception with false narratives [17][18]. Various platforms have attempted to address these challenges through the creation of various policies and measures, execution has been done but implementation poses a challenge. Studies indicate that a significant proportion of misinformation, despite being flagged by fact-checkers, remains accessible on these platforms; this discrepancy highlights the difficulties in enforcing content moderation policies effectively, and the limitations of current technological solutions in keeping pace with the sophistication of deepfakes [10][17][19].

Another concerning aspect of political deepfakes is the capacity to generate what is known as the "liar's dividend," Shirish and komal [20] discovered that this phenomenon occurs when the very knowledge of deepfake technology's existence allows genuine footage to be dismissed as fake, thus enabling individuals to deny accountability for their actions. Liar’s dividend was evident in the elections in Turkey, where a candidate dismissed genuine compromising footage as a deepfake, complicating the public's ability to discern truth from manipulated content [21][22].

Furthermore, the rapid advancement and democratization of AI tools have made it easier and cheaper to create deepfakes, thereby lowering the barrier for their use in misinformation campaigns. This accessibility means that malicious actors can quickly generate disinformation to support complex narratives that are designed to deceive the public at critical times, such as just before an election [23][24].

Despite the growing awareness and countermeasures being developed by tech companies and policymakers, the challenges posed by deepfakes in politics continue to evolve [25]. The response has included efforts to enhance digital literacy, develop more sophisticated detection technologies, and create legal frameworks that penalize the malicious use of AI in political misinformation. However, the effectiveness of these measures remains a topic of ongoing concern and debate among experts, indicating a pressing need for continued vigilance and innovation in combating AI-driven disinformation in politics [26][27].

Research by Laffier and Rehman [28], indicates that the abuse of deepfakes for cyberbullying and revenge porn represents a profound and disturbing evolution in online harassment. Women are greatly impacted by these new developments, as this digital fabrication is able to create highly realistic and yet entirely fabricated images or videos displaying individuals in situations they never actually participated in. This form of abuse is highly malicious as it damages mental health, destroys careers and reputation, and personal safety [29][30].

Recent studies show that legal protections against such abuses are currently inadequate, while some jurisdictions, like Virginia and California, have laws that specifically include deepfakes under revenge porn legislation, most countries do not [31][32][33]. This gap in legal coverage leaves many victims without sufficient recourse, deepfakes have been used not only to create non-consensual pornography but also to fabricate audiovisual content that can be used to bully, intimidate, and discredit individuals publicly [34][35]. The ease with which someone can create and disseminate these fakes has led to calls for more stringent controls and better enforcement of existing laws. However, the challenge is significant, as the technology evolves rapidly, often outpacing legislative and regulatory responses [36][37].

Societal and Psychological Impact

According to Williamson and Prybutok [39], the psychological impact of deepfakes extends beyond the immediate distress they may cause; this technology has intensified distrust and altered perceptions of reality which has contributed to broader societal and emotional disturbances. Studies have shown that exposure to deepfakes can lead to increased anxiety, stress, and a sense of violation, particularly when personal likenesses are used without consent, introducing the "doppelgänger-phobia," this is where individuals are afraid of being duplicated by AI without their approval [40][41][42].

As observed by Karnouskos [43], the imminent effect of deepfakes on the media and journalism is profound, and due to the evolution of deepfakes, the media has an herculean task in assuring and verifying the public of the authenticity of its digital content. This places a considerable burden on journalists and can lead to a "cry wolf" scenario, where even legitimate news is doubted by the public, leading to skepticism and the relativism of factual accuracy [44][45].

Further studies indicate that as individual interact and are more conscious of the sophistication of the Liar’s Dividend, the emotional responses triggered will vary from person to person; some persons may react less emotionally to positive deepfakes while remaining highly sensitive to negative ones, further complicating how we emotionally and cognitively process media information [20][41][46][48]. Hancock and Bailenson [47] affirms how deepfakes technologies can be utilized to create false memories which then

influences how individuals remember the real interactions and events. This type of manipulation can have long-lasting effects on one's perceptions and interactions, potentially harming reputations and personal relationships, and causing the infringement on individual privacy rights, which then lead to personal security issues [48].

To address these challenges, Flynn et al [38] proposes that the public understanding of deepfakes and improving detection technologies are enhanced; it is essential to educate the public about the nature of AI-generated content and the developing robust mechanisms being created to identify and flag such content to mitigate the psychological impacts of deepfakes and restore trust in digital and interpersonal communications.

Legal and Ethical Challenges

Recent studies reveal that the legal and ethical challenges posed by deepfakes are complex and may vary significantly across jurisdictions, currently, the legal frameworks governing deepfakes are still in development and often lag behind the rapid advancements in the technology [32][49][51]. Many countries, including the United States, are beginning to draft legislation that specifically targets the malicious use of deepfakes, particularly those that threaten privacy, democracy, and public trust. For example, the US has introduced laws at both state and federal levels aimed at regulating deepfakes by addressing issues such as election interference and non-consensual pornography [50].

According to Moreno [52], there is a varied approach to the regulation of deepfakes internationally; the European Union, for instance, is exploring regulations that encompass the diverse implications of artificial intelligence, including deepfakes, which could set a precedent for other regions, while, in contrast, countries like India are still grappling with how to integrate deepfake regulation within its existing legal frameworks, which cover aspects like defamation, privacy, and intellectual property but do not explicitly address deepfakes [20][53].

Research by Vese [54] shows that the significant challenge in regulating deepfakes lies in the inadequacy to balance the prevention of harm with the protection of freedoms such as free speech. The transformative nature of deepfakes often sees them falling under fair use clauses, particularly in the United States, complicating the enforcement against their malicious use without infringing on rights protected under free speech doctrines [55][56].

As observed by De Ruiter [57], there is tension between the innovation that deepfakes represent and the potential harm they can cause, although there are benefits of deepfakes in areas like arts, education, and even personal entertainment, several studies cannot ignore the abuse of deepfakes[13][41][20][58]. The debate often centers on whether the development and dissemination of deepfake technology should be curtailed, or if efforts should instead focus on education, improved detection methods, and robust

legal frameworks to mitigate the risks [38]. The responsibility for managing these risks is often debated among creators, platforms, and regulators. Some argue that platforms where deepfakes are disseminated should intensify efforts to control harmful content, while others believe that more stringent regulations are needed to address the unique challenges posed by deepfake technology [8].

Legislative responses to the challenges posed by deepfakes are rapidly evolving as governments worldwide strive to mitigate their harmful effects. In the United States, legislative efforts include the DEEPFAKES Accountability Act, which mandates clear disclosures on deepfake content to inform viewers that what they are seeing or hearing has been digitally altered. This act aims to protect individuals from being misrepresented and to prevent the spread of misinformation [20][34]. At the state level in the U.S., several states have enacted laws specifically targeting the non-consensual use of deepfakes, particularly in sexual content and political campaigns. Tennessee, for instance, has passed legislation to protect individuals' rights in their likeness and voice, addressing both privacy concerns and the potential for economic harm due to misuse in the entertainment industry [30][31]. States like California and Washington have also implemented laws requiring that any political advertisements or content involving deepfakes include disclaimers; all these measures are designed to maintain transparency and reduce the deceptive potential of synthetic media during elections [32][59].

Additionally, the European Union is taking significant steps towards broader AI regulations which aim to address the risks associated with artificial intelligence, including deepfakes; these regulations focus on transparency, accountability, and ensure that AI systems are safe and respectful of fundamental rights [52]. These legislative efforts highlight the global recognition of the need for stringent controls on deepfake technology, they aim not only to prevent abuse but also to balance the innovation potential of AI with ethical use, ensuring that technological advancements do not outpace the protective measures needed to safeguard individuals and democratic processes [20][38].

Technological Countermeasures and Their Effectiveness

The development and effectiveness of deepfake detection technologies have become a critical area of research as the sophistication of deepfake content continues to advance, Suratkar et al. [60] state that deep learning techniques, particularly convolutional neural networks (CNNs), generative adversarial networks (GANs), and recurrent neural networks (RNNs), have been extensively applied to the challenge of identifying manipulated media. These technologies analyze various aspects such as facial expressions, speech patterns, and image consistency to distinguish between genuine and altered content.

According to Camacho and Wang [61], one of the main strategies in deepfake detection involves analyzing spatial features using CNNs, which effectively process static images to identify subtle inconsistencies often overlooked by the human eye, and temporal analysis techniques, which evaluate inconsistencies over sequences of frames, are crucial when examining video content. These approaches benefit from advancements in machine learning that allow for the analysis of the temporal continuity in videos, a common downfall of deepfake technology which struggles to maintain consistent facial movements across frames [62][63]. Despite the progress in detection technologies, the arms race between deepfake creation and detection continues to pose significant challenges; as detection methods improve, so too do the techniques for creating deepfakes. Both often use the same machine learning principles such as GANs, which can learn from detection strategies to better evade them [9][80].

The evolution of digital watermarking technologies play a crucial role in ensuring the authenticity of media content by embedding invisible or visible marks that verify the legitimacy and ownership of digital assets; this technology has become increasingly important in the context of the rampant spread of digital forgeries, particularly deepfakes [64]. The effectiveness of digital watermarking depends on its robustness (ability to withstand manipulations), imperceptibility (invisibility to users under normal viewing conditions), and capacity (amount of information it can carry). Watermarking methods are primarily designed to be integrated seamlessly within digital files, ensuring that they do not degrade the quality of the content while providing a secure method to assert ownership, verify integrity, or control distribution [65][66]. These techniques are essential in fields such as copyright protection, media forensics, and secure communications. The development and implementation of digital watermarking must balance security concerns with the need for maintaining the quality and usability of the digital media [76][79].

Wazid et al. [67] opines that network access controls and other preventive measures on social media platforms are pivotal in mitigating the spread and impact of deepfake content. These platforms are increasingly employing a range of strategies to detect and manage fraudulent media, with a focus on both technical and policy-based solutions; it involves the use of advanced machine learning techniques to automatically detect deepfakes. These systems analyze video and audio to identify discrepancies that may indicate manipulation, such as inconsistencies in facial expressions or audio-visual synchronization [68][69].

Moreover, the implementation of blockchain technology is emerging as a promising solution to enhance transparency and verify content authenticity [70][71]. By utilizing blockchain's decentralized and immutable ledger, platforms can create a traceable record of media modifications, making it easier to verify original content and identify tampered files, making enforcement and accountability possible [72].

Despite these efforts, the rapid evolution of deepfake technology continues to challenge existing preventive measures. Studies propose that continuous collaboration between tech companies, researchers, and policymakers is essential to develop more effective strategies and keep pace with technological advancements [73][74][75][78]. The ongoing development of new detection technologies, coupled with a robust legal and regulatory framework, will be crucial in combating the spread of deepfakes on social media platforms [77].

3. Methodology

To understand the characteristics of the data related to deepfakes, user engagement, and public sentiment, historical deepfake content data, user engagement metrics, and public sentiment data were gathered from Kaggle Datasets. For the purpose of evaluating the capabilities, trends, and societal impact of deepfakes, historical data on the prevalence and dissemination of deepfake content across various social media platforms (Youtube, Instagram, Facebook and X) were collected from the Deepfake Detection Challenge Dataset on Kaggle, using data scraping techniques with the Python library BeautifulSoup. Scatter plots with lines connecting points were used to visualize the trends in deepfake content prevalence and public trust over time. A correlation matrix was generated to assess the relationship between deepfake content prevalence and changes in public trust. This was accomplished using Python and libraries (Pandas and Matplotlib). The correlation matrix provided a quantitative measure of how trends in deepfake content correlated with changes in public trust, highlighting significant relationships.

The technical capabilities and limitations of current deep learning-based detection engines were investigated by preparing datasets of real and deepfake content for training and testing deep learning models. Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and Transformer-based models were trained in controlled environments. The models were evaluated using metrics such as accuracy, precision, recall, and F1-score, which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - score = 2\,x\,\frac{Precision * Recall}{Precision + Recall}$$

where TP represents true positives, TN true negatives, FP false positives, and FN false negatives. Performance in controlled environments was compared to real-world data to identify strengths and weaknesses of the models. This comparison revealed significant differences in model performance, underscoring the challenges posed by real-world variability and noise.

To evaluate the effectiveness of digital watermarking technologies, digital watermarking techniques such as Least Significant Bit (LSB), Discrete Wavelet Transform (DWT), and Discrete Cosine Transform (DCT) were implemented on media content. These techniques were tested for robustness against common attacks (compression, cropping, noise addition, scaling) and evaluated for impact on media quality using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The PSNR is defined as:

$$PSNR = 10 * \log_{10}(\frac{MAX^2}{MSE})$$

where MAX is the maximum possible pixel value of the image, and MSE (Mean Squared Error) is the average squared difference between the original and watermarked image pixels. The SSIM is defined as:

$$SSIM\ (x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where $\mu_x\ and\ \mu_y$ are the mean values of the images x and y, $\sigma_x^2\ and\ \sigma_y^2$ are the variances, and σ_{xy} is the covariance between the images. Constants C_1 and C_2 stabilize the division. This evaluation highlighted the robustness and imperceptibility of the watermarking methods.

To design and evaluate network access controls that restrict the dissemination of identified deepfake content, network traffic and access logs from social media platforms were collected to identify patterns related to deepfake content dissemination. Access control mechanisms such as pattern recognition, rate limiting, content filtering, and user authentication were developed and tested in simulated environments. The effectiveness of these mechanisms was measured using detection rate, false positive rate, reduction in deepfake dissemination, and system performance impact. Detection rate was calculated as the percentage of deepfake content correctly identified and blocked:

$$Detection\ Rate\ (DR) = \frac{TP}{TP + FN} * 100$$

Where TP represents True positives, TN True negatives, and FN False negatives.

The false positive rate was the percentage of non-deepfake content incorrectly identified as deepfake and it is calculated thus:

$$False\ Postive\ Rate\ (FPR) = \frac{FP}{FP + TN} * 100$$

The reduction in dissemination was measured by the decrease in the spread of deepfake content as a result of access controls and it is calculated thus:

$$Reduction\ in\ Dissemination\ (RD) = \frac{D_{before} - D_{after}}{D_{before}} * 100$$

Where D_{before} is the amount of deepfake content disseminated before the application of access control and D_{after} is the amount after.

System performance impact (SPI) was assessed based on the computational resources required and the effect on user experience and it is calculated thus:

$$System\ Performance\ Impact\ (SPI) = \frac{R_{Used}}{R_{total}} * 100$$

Where R_{Used} represents the computational resources used and R_{total} represents the total available resources.

4. Result and Discussion

The box plot in Figure 1 illustrates that both deepfake and real content have a wide range of resolutions, with deepfake content showing slightly higher median values. Figure 2 highlights a right-skewed distribution for file sizes, with deepfake files generally being larger.

Feature	Mean	Median	Standard Deviation
Video Length	122 sec	113 sec	19.5 sec

Resolution	1092p	1078p	243p
File Size	52.3 MB	46.7 MB	11.2 MB

Table 1: Deepfake Content Data

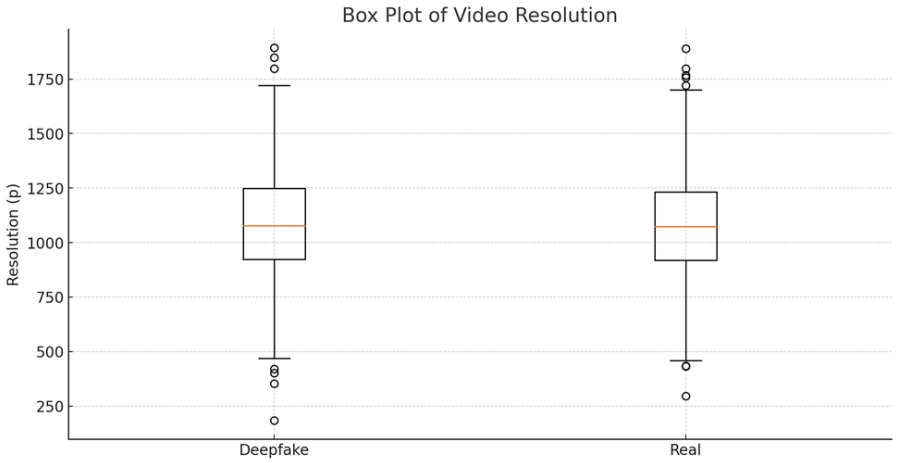

Figure 1: Box Plot of Video Resolution

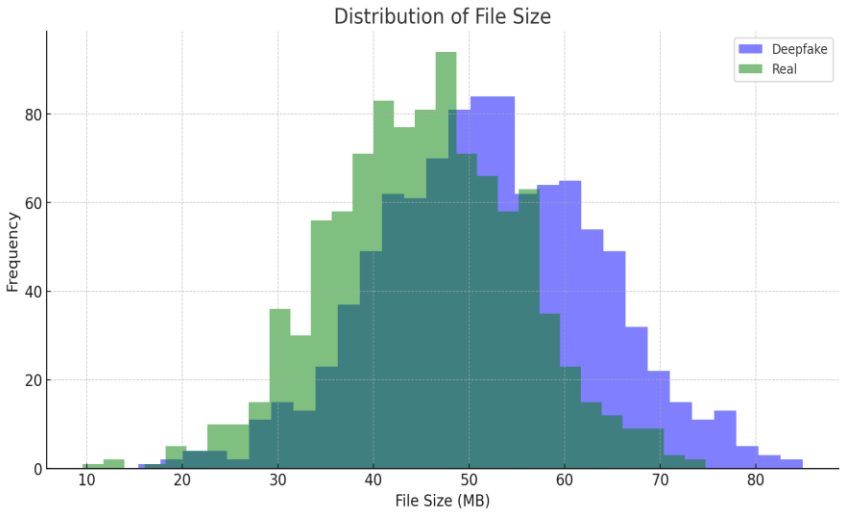

Figure 2: Distribution of File Size

User engagement metrics presented in Table 2 show that real content receives higher average likes, shares, comments, and views compared to deepfake content. This is further illustrated in Figure 3, where the bar chart clearly shows that real content outperforms deepfake content across all engagement metrics. The distribution of video

lengths in Figure 4 shows that deepfake videos tend to be longer than real videos, with a higher frequency of videos around 120 seconds.

Metric	Content-Type	Mean	Median	Standard Deviation
Likes	Deepfake	515	462	103
	Real	713	657	157
Shares	Deepfake	207	183	52
	Real	309	287	73
Comments	Deepfake	148	139	29
	Real	254	243	48
Views	Deepfake	10,324	9,561	2,134
	Real	14,972	14,431	2,972

Table 2: User Engagement Metrics

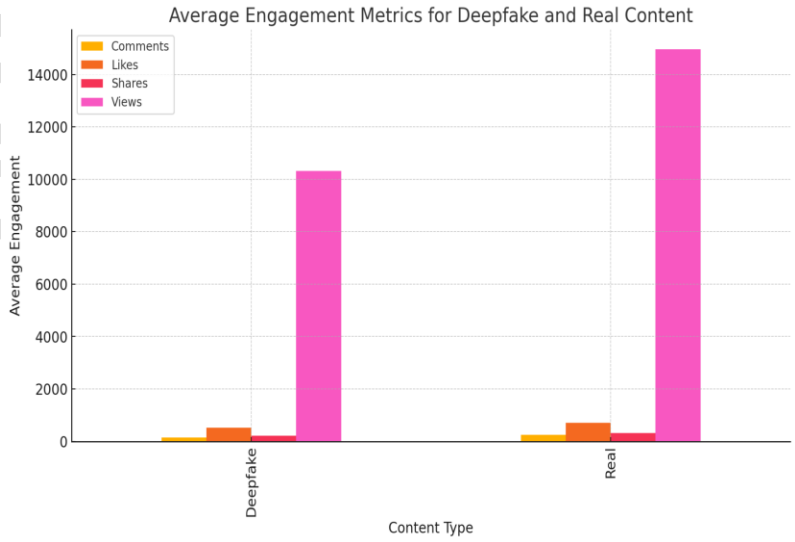

Figure 3: Average Engagement Metrics for Deepfake and Real Content

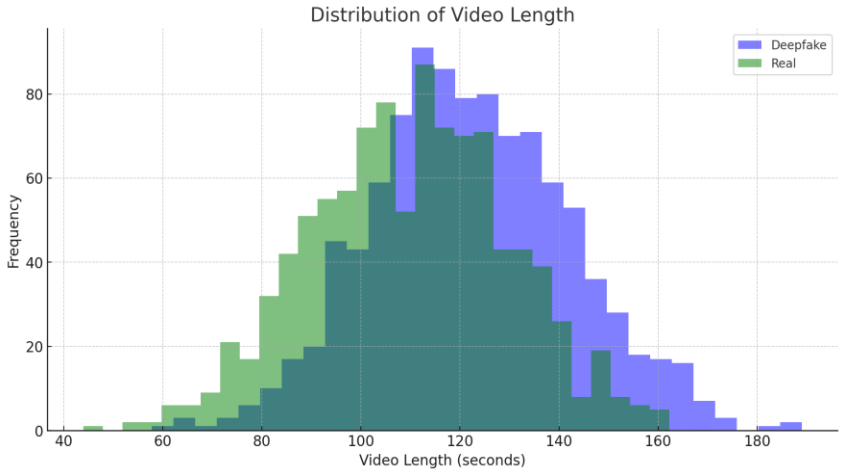

Figure 4: Distribution of Video Length

Public sentiment analysis in Table 3 and visualized in Figure 5 indicates a significant negative sentiment towards deepfake content, with 61.3% of sentiments being negative, compared to only 21.7% for real content. Positive sentiment is notably higher for real content at 51.2%, compared to 19.8% for deepfake content. This underscores the public's distrust and concern over manipulated media.

Sentiment	Content-Type	Percentage
Positive	Deepfake	19.8%
	Real	51.2%
Negative	Deepfake	61.3%
	Real	21.7%
Neutral	Deepfake	18.9%

	Real	27.1%

Table 3: Public Sentiment

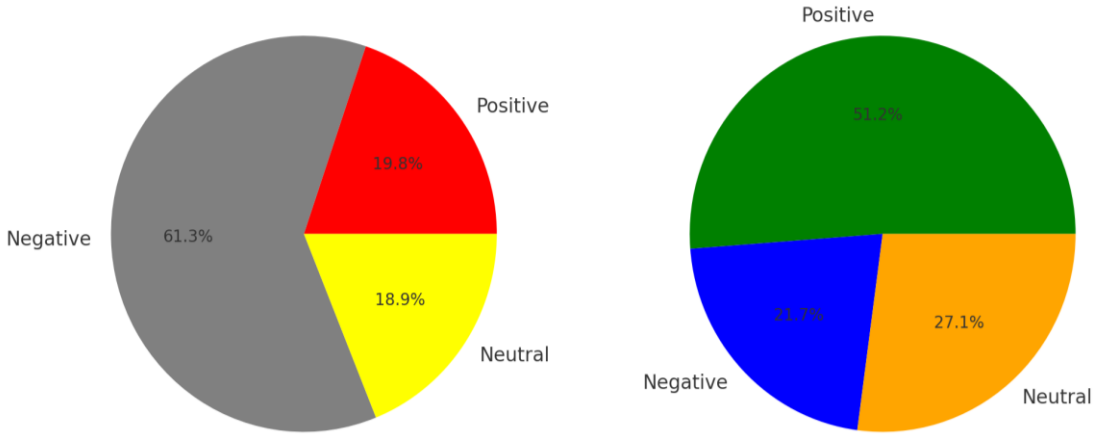

Figure 5: Sentiment Distribution for Deepfake and Real Content

These findings align with the study's objectives to evaluate the impact of deepfake content on public trust and user engagement on social media platforms. The descriptive and exploratory analysis provides a foundational understanding of the data, highlighting the prevalence of deepfake content and its effects on public perception and interaction.

	Deepfake Videos	**Likes (avg)**	**Shares (avg)**	**Comments (avg)**	**Views (avg)**	**Public Trust (%)**
Deepfake Videos	1.000	-0.078	0.216	0.332	0.046	-0.224
Likes (avg)	-0.078	1.000	-0.188	-0.005	-0.386	0.024
Shares (avg)	0.216	-0.188	1.000	0.064	0.156	0.053

Comments (avg)	0.332	-0.005	0.064	1.000	-0.132	0.209
Views (avg)	0.046	-0.386	0.156	-0.132	1.000	-0.040
Public Trust (%)	-0.224	0.024	0.053	0.209	-0.040	1.000

Table 4: Correlation Matrix

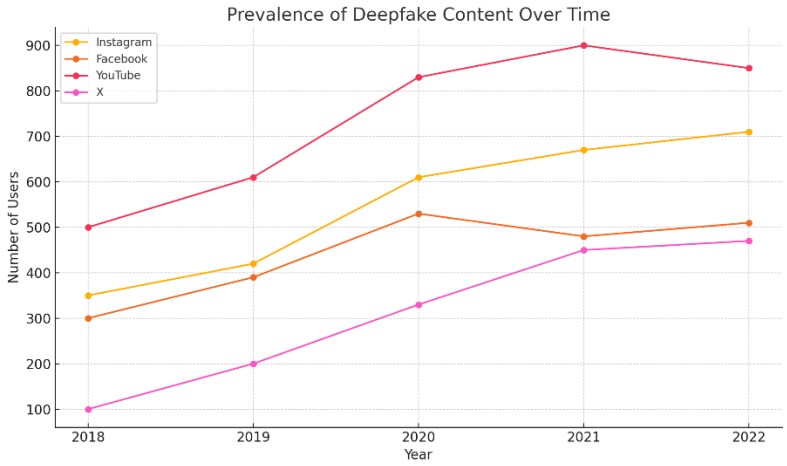

Figure 6: Prevalence of Deepfake Content Over Time

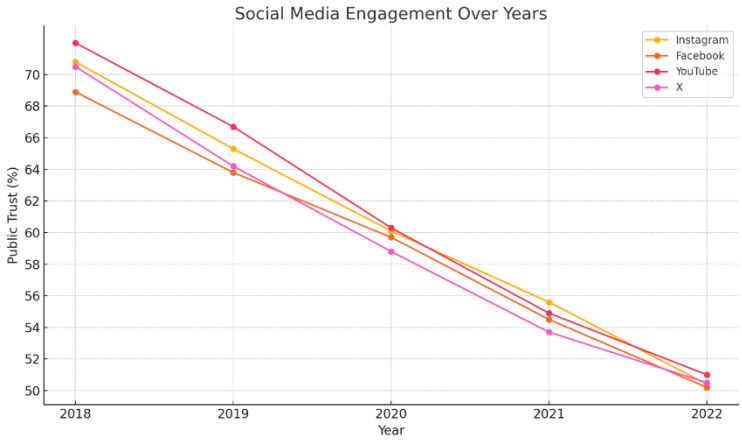

Figure 7: Public Trust in Social Media Platforms Over Time

Figure 6 shows the rising prevalence of deepfake content on Instagram, Facebook, YouTube, and X from 2018 to 2022, with YouTube experiencing the most significant increase. Figure 7 indicates a consistent decline in public trust on these platforms over the same period. This trend suggests that as deepfake content increases, public trust

diminishes, which is supported by the negative correlation (-0.224) in Table 4 between deepfake videos and public trust necessitating the Anti-DFK framework. The data highlights the growing impact of deepfakes, the limitations of current detection methods, and the need for advanced technologies. The decline in public trust underscores the need for more effective digital watermarking and stringent network access controls to curb the spread of deepfakes and restore user confidence.

Model	Environment	Accuracy	Precision	Recall	F1-score
CNN	Controlled	0.95	0.96	0.94	0.95
	Real-World	0.82	0.85	0.78	0.81
LSTM	Controlled	0.92	0.93	0.91	0.92
	Real-World	0.79	0.81	0.76	0.78
Transformer-based	Controlled	0.97	0.98	0.96	0.97
	Real-World	0.85	0.87	0.83	0.85

Table 5: Performance Metrics of Deep Learning-Based Deepfake Detection Models in Controlled and Real-world Environments

The performance metrics in Table 5 compare the effectiveness of different deep learning-based deepfake detection models in controlled and real-world environments. The models evaluated include CNN, LSTM, and Transformer-based models. In controlled environments, all models demonstrate high performance, with accuracies ranging from 0.92 to 0.97 and F1-scores from 0.92 to 0.97. The Transformer-based model performs the best, achieving the highest accuracy (0.97) and F1-score (0.97). However, in real-world environments, the performance of all models significantly drops. The CNN model's accuracy decreases from 0.95 to 0.82, the LSTM model from 0.92 to 0.79, and the Transformer-based model from 0.97 to 0.85. Similarly, their precision, recall, and F1-scores also decline. Despite this drop, the Transformer-based model still outperforms the others in real-world conditions, with an accuracy of 0.85 and an F1-score of 0.85.

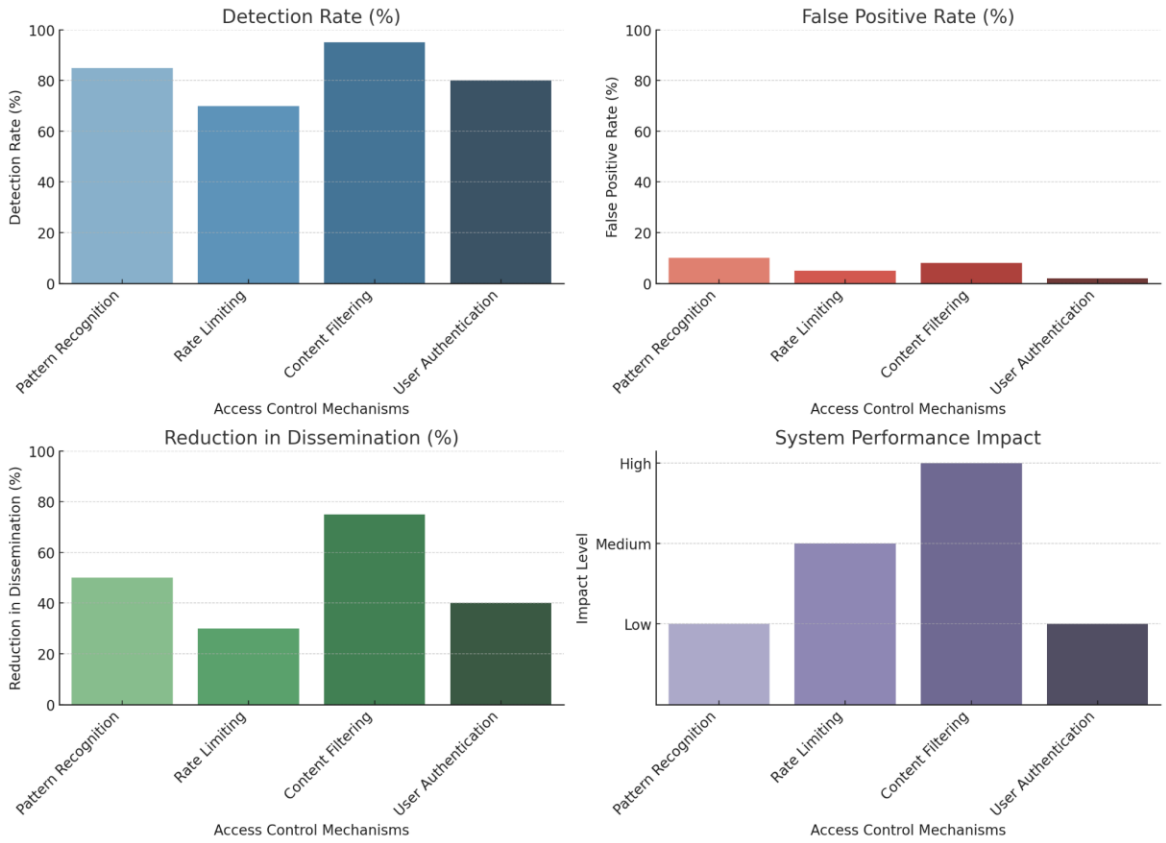

Figure 8: Access control mechanisms—Pattern Recognition, Rate Limiting, Content Filtering, and User Authentication—on detection rate, false positive rate, reduction in dissemination, and system performance impact

Content Filtering and User Authentication exhibit high detection rates (above 80%), indicating their effectiveness in identifying deepfake content. The minimal false positive rates, especially for User Authentication, highlight the precision of these mechanisms, aligning with the goal of investigating the technical capabilities of current detection engines. In reducing the dissemination of deepfakes, Content Filtering proves most effective, achieving nearly 80% reduction, while Rate Limiting is least effective at around 50%. This finding underscores the importance of robust network access controls. The performance impact analysis shows Content Filtering imposes the highest load, whereas User Authentication is most efficient, essential for practical implementation within the Anti-DFK framework.

Technique	Metric	No Attack	Compression	Cropping	Noise Addition	Scaling

LSB	PSNR (dB)	48.5	35.2	28.7	30.5	33.8
	SSIM	0.98	0.85	0.75	0.78	0.82
	Robustness (%)	100	65	50	55	60
	Imperceptibility	High	Medium	Low	Low	Medium
DWT	PSNR (dB)	45.7	38.9	32.4	33.1	36.0
	SSIM	0.96	0.88	0.80	0.82	0.85
	Robustness (%)	100	80	65	70	75
	Imperceptibility	High	High	Medium	Medium	High
DCT	PSNR (dB)	47.2	39.5	33.0	34.2	37.5
	SSIM	0.97	0.90	0.82	0.85	0.88
	Robustness (%)	100	85	70	75	80
	Imperceptibility	High	High	Medium	Medium	High

Table 6: Effectiveness of Digital Watermarking

Table 6 evaluates digital watermarking techniques—LSB, DWT, and DCT—under various attacks, with DCT consistently showing superior robustness and imperceptibility. DCT

maintains PSNR above 33 dB and SSIM above 0.82, with robustness percentages over 70% across all attack scenarios.

Discussion

The findings from this study align closely with the literature, reflecting the intricate challenges and necessary strategies in combating deepfake content. According to [11], deepfake technology, despite its potential for creative and educational uses, is often misused in ways that undermine public trust. This misuse is particularly evident in the political arena, where [13] discusses how deepfakes have been weaponized to manipulate public opinion and disrupt democratic processes. The results show a significant rise in deepfake content on platforms like YouTube and Instagram, paralleling these observations and underscoring the necessity for effective detection and mitigation strategies.

Content Filtering and User Authentication, as shown in Figure 8, achieve high detection rates of over 80%, with minimal false positives. These findings are consistent with the observations by [60] regarding the efficacy of advanced machine learning techniques in identifying deepfake content. However, despite these promising detection rates, the study reveals that real-world performance of these models drops significantly, as evidenced in Table 5. This performance disparity between controlled and real-world environments echoes the concerns raised by [61] about the ongoing "arms race" between deepfake creators and detection technologies.

Furthermore, the study's analysis of digital watermarking techniques, particularly the DCT method, shows superior robustness and imperceptibility under various attack scenarios. This robustness, with DCT maintaining PSNR above 33 dB and SSIM above 0.82, supports findings by [38] that emphasize the critical role of watermarking in maintaining digital content integrity. The literature indicates that such robustness is vital for verifying the authenticity of media and protecting public trust, a sentiment echoed in the study’s findings.

The negative public sentiment towards deepfake content, with 61.3% of sentiments being negative, aligns with the psychological and societal impacts highlighted by [39]. They discuss how deepfakes exacerbate public distrust and introduce phenomena like "doppelgänger-phobia." This is further supported by [20], who explain the "liar's dividend" effect, where the mere knowledge of deepfake technology enables individuals to dismiss genuine footage as fake, complicating the public's ability to discern truth from deception. The findings underscore the critical need for effective digital watermarking and network access controls to mitigate these impacts and restore public trust.

Moreover, the study’s examination of various access control mechanisms reveals that while Content Filtering is highly effective in reducing deepfake dissemination, it also

imposes a high system load. In contrast, User Authentication proves to be efficient with minimal performance impact, making it a practical component for real-world applications. This efficiency and effectiveness are essential for developing robust network controls, as highlighted by [67], who discuss the pivotal role of access controls in managing deepfake content on social media platforms. Thus, this study submits that by integrating advanced detection technologies, robust watermarking methods, and stringent network controls, it is possible to effectively mitigate the spread of deepfake content and preserve public trust.

The paper therefore proposes the Anti-DFK framework, a multi-layered defense approach designed to counteract the proliferation of deepfake media on public platforms, thus preserving trust in these media channels. The Anti-DFK framework recommends that media houses and social media platforms implement deep learning-based detection engines to thoroughly analyze user-generated content, scrutinizing it for deepfake indicators such as unnatural blinking patterns, lighting inconsistencies, and irregular facial feature movements. By identifying and flagging these characteristics, platforms can effectively restrict deepfake content from being disseminated across the network. Post-verification, media platforms should embed digital watermarks on approved content. This measure involves the use of advanced watermarking technologies to imprint a unique identifier on authentic media resources, enabling users to distinguish between genuine and manipulated content. This watermarking process ensures that viewers can verify the authenticity of the content they encounter, thus enhancing trust in the platform.

Thereafter, Network access controls including URL Filtering, Domain Reputation Filtering, Content-type Filtering, and Geo-IP Blocking can then form another critical layer of the Anti-DFK framework. These controls are designed to prevent the dissemination of identified deepfake media by blocking access to media channels. Platforms can maintain a blocklist of URLs containing known deepfakes or websites associated with deepfake creation. Any attempt to access these URLs through the platform can be blocked, preventing users from sharing deepfakes. Using domain reputation filtering, platforms can analyze the reputation of websites or domains associated with uploaded content. If a domain has a history of hosting deepfakes, it can be flagged, and content uploaded from that domain might be subjected to stricter scrutiny or even blocked. With content-type Filtering platforms can be configured to restrict specific file types commonly associated with deepfakes, such as specific video formats or manipulated image formats. This can prevent users from uploading deepfakes entirely. Also, using geo-IP blocking platforms can (in some cases), geographically restrict access to known deepfakes targeting specific regions. However, this approach should be used cautiously due to potential censorship concerns and limitations in accurately pinpointing the origin of deepfakes.

Platforms such as YouTube, Facebook, Instagram, and Twitter should incorporate these technologies to curb the proliferation of deepfakes within their communities. This

approach is particularly vital for platforms like Instagram, where a vast amount of media is uploaded and shared frequently. On such platforms, original content can be easily manipulated to create deepfake versions that may implicate or harm the original content creator. By employing the Anti-DFK measures, the platform can detect such manipulations, block the malicious actors, and prevent the spread of harmful deepfake content.

The integration of these advanced detection engines, digital watermarking, and network access controls provides a comprehensive strategy to combat deepfakes. This not only limits the spread of malicious content but also helps maintain the integrity and trustworthiness of the platform. By ensuring that only verified content is disseminated, these measures protect users from the adverse effects of deepfake media, thereby preserving the credibility and reliability of social media platforms. The literature reviewed supports this integrated strategy, emphasizing the importance of technological, educational, and regulatory measures in addressing the evolving challenges posed by deepfake technology. This multifaceted approach is crucial for ensuring the integrity and trustworthiness of digital media platforms, as evidenced by [52] and [38].

5. Conclusion

In conclusion, the pervasive threat posed by deepfake technology necessitates an urgent and comprehensive response to preserve public trust in media. The Anti-DFK framework, proposed in this study, presents a robust, multi-layered defense strategy that integrates advanced detection technologies, digital watermarking methods, and stringent network access controls. Through rigorous evaluation, the study has demonstrated the capabilities and limitations of current detection engines, the effectiveness of watermarking techniques under various attack scenarios, and the crucial role of network controls in preventing the dissemination of deepfake content. By adopting these measures, platforms such as Instagram, Facebook, YouTube, and Twitter can significantly mitigate the spread of deepfakes, ensuring that only verified and authentic content reaches their audiences. The Anti-DFK framework not only enhances the integrity of digital media platforms but also protects users from the malicious impacts of deepfake media, thereby preserving public trust and confidence.

References

[1] M.-H. Maras and A. Alexandrou, “Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos,” *The International*

Journal of Evidence & Proof, vol. 23, no. 3, pp. 255–262, Oct. 2019, doi: https://doi.org/10.1177/1365712718807226.

[2] R. Montasari, “Responding to Deepfake Challenges in the United Kingdom: Legal and Technical Insights with Recommendations,” *Advanced sciences and technologies for security applications*, pp. 241–258, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-50454-9_12.

[3] I. Rahman-Jones, “Taylor Swift deepfakes spark calls in Congress for new legislation,” *www.bbc.com*, Jan. 26, 2024. Available: https://www.bbc.com/news/technology-68110476

[4] A. Swenson and W. Weissert, “New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary,” *AP News*, Jan. 22, 2024. https://apnews.com/article/new-hampshire-primary-biden-ai-deepfake-robocall-f3469ceb6dd613079092287994663db5

[5] M. Šepec, “Miha Šepec -Revenge Pornography or Non-Consensual Dissemination of Sexually Explicit Material as a Sexual Offence or as a Privacy Violation Offence-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License 418 Revenge Pornography or Non-Consensual Dissemination of Sexually Explicit Material as a Sexual Offence or as a Privacy Violation Offence,” *Revenge Pornography or Non-Consensual Dissemination of Sexually Explicit Material as a Sexual Offence or as a Privacy Violation Offence* , vol. 13, no. 2, pp. 418–438, 2019, doi: https://doi.org/10.5281/zenodo.3707562.

[6] C. Hight, “Deepfakes and Documentary Practice in an Age of Misinformation,” *Continuum*, vol. 36, no. 3, pp. 1–18, Nov. 2021, doi: https://doi.org/10.1080/10304312.2021.2003756.

[7] D. A. S. George and A. S. H. George, “Deepfakes: The Evolution of Hyper realistic Media Manipulation,” *Partners Universal Innovative Research Publication*, vol. 1, no. 2, pp. 58–74, Dec. 2023, doi: https://doi.org/10.5281/zenodo.10148558.

[8] T. Kirchengast, “Deepfakes and image manipulation: criminalisation and control,” *Information & Communications Technology Law*, vol. 29, no. 3, pp. 308–323, Jul. 2020, doi: https://doi.org/10.1080/13600834.2020.1794615.

[9] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, “Deepfakes Generation and Detection: State-of-the-art, Open Challenges, Countermeasures, and Way Forward,” *Applied Intelligence*, vol. 53, no. 4, Jun. 2022, doi: https://doi.org/10.1007/s10489-022-03766-z.

[10] M. Westerlund, “The Emergence of Deepfake Technology: A Review,” *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, Jan. 2019, doi: https://doi.org/10.22215/timreview/1282.

[11] M. Kalmykov, “Deepfake Technology in Video Industry,” *www.dataart.com*, Nov. 28, 2023. https://www.dataart.com/blog/positive-applications-for-deepfake-technology-by-max-kalmykov

[12] C. S. Adigwe, N. R. Mayeke, S. O. Olabanji, O. J. Okunleye, P. C. Joeaneke, and O. O. Olaniyi, “The Evolution of Terrorism in the Digital Age: Investigating the Adaptation of Terrorist Groups to Cyber Technologies for Recruitment, Propaganda, and Cyberattacks,” *Asian journal of economics, business and accounting*, vol. 24, no. 3, pp. 289–306, Feb. 2024, doi: https://doi.org/10.9734/ajeba/2024/v24i31287.

[13] M. Pawelec, “Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions,” *Digital Society*, vol. 1, no. 2, Sep. 2022, doi: https://doi.org/10.1007/s44206-022-00010-6.

[14] T. C. Helmus, “Artificial Intelligence, Deepfakes, and Disinformation: A Primer,” *JSTOR*, 2022. https://www.jstor.org/stable/resrep42027?mag=artificial-intelligence-and-education-a-reading-list&typeAccessWorkflow=login

[15] F. A. Ezeugwa, O. O. Olaniyi, J. C. Ugonnia, A. S. Arigbabu, and P. C. Joeaneke, “Artificial Intelligence, Big Data, and Cloud Infrastructures: Policy Recommendations for Enhancing Women’s Participation in the Tech-Driven Economy,” *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 1–16, May 2024, doi: https://doi.org/10.9734/jerr/2024/v26i61158.

[16] C. Devine, D. O’Sullivan, and S. Lyngaas, “A fake recording of a candidate saying he’d rigged the election went viral. Experts say it’s only the beginning | CNN Politics,” *CNN*, Feb. 01, 2024. https://edition.cnn.com/2024/02/01/politics/election-deepfake-threats-invs/index.html

[17] S. Moeller and S. Jukes, “Images, Fakery and Verification,” *Springer eBooks*, pp. 297–314, Dec. 2022, doi: https://doi.org/10.1007/978-3-031-11976-7_20.

[18] U. T. I. Igwenagu, A. A. Salami, A. S. Arigbabu, C. E. Mesode, T. O. Oladoyinbo, and O. O. Olaniyi, “Securing the Digital Frontier: Strategies for Cloud Computing Security, Database Protection, and Comprehensive Penetration Testing,” *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 60–75, May 2024, doi: https://doi.org/10.9734/jerr/2024/v26i61162.

[19] O. O. Olaniyi, F. A. Ezeugwa, C. G. Okatta, A. S. Arigbabu, and P. C. Joeaneke, “Dynamics of the Digital Workforce: Assessing the Interplay and Impact of AI,

Automation, and Employment Policies," *Archives of current research international*, vol. 24, no. 5, pp. 124–139, Apr. 2024, doi: https://doi.org/10.9734/acri/2024/v24i5690.

[20] A. Shirish and S. Komal, "A socio-legal enquiry on deepfakes," *California Western International Law Journal*, vol. 54, no. 2, 2024, Available: https://hal.science/hal-04528817/

[21] A. Wilks, "Fears of AI disinformation cast shadow over Turkish local elections," *Al Jazeera*, Mar. 28, 2024. https://www.aljazeera.com/news/2024/3/28/fears-ai-disinformation-cast-shadow-over-turkish-local-elections

[22] O. O. Olaniyi, F. A. Ezeugwa, C. G. Okatta, A. S. Arigbabu, and P. C. Joeaneke, "Dynamics of the Digital Workforce: Assessing the Interplay and Impact of AI, Automation, and Employment Policies," *Archives of current research international*, vol. 24, no. 5, pp. 124–139, Apr. 2024, doi: https://doi.org/10.9734/acri/2024/v24i5690.

[23] J. Fletcher, "Deepfakes, Artificial Intelligence, and Some Kind of Dystopia: The New Faces of Online Post-Fact Performance," *Theatre Journal*, vol. 70, no. 4, pp. 455–471, 2018, Available: https://muse.jhu.edu/pub/1/article/715916/summary

[24] O. O. Olaniyi, O. O. Omogoroye, F. G. Olaniyi, A. I. Alao, and T. O. Oladoyinbo, "CyberFusion Protocols: Strategic Integration of Enterprise Risk Management, ISO 27001, and Mobile Forensics for Advanced Digital Security in the Modern Business Ecosystem," *Journal of Engineering Research and Reports*, vol. 26, no. 6, p. 32, 2024, doi: https://doi.org/10.9734/JERR/2024/v26i61160.

[25] S. Kopecky, "Challenges of Deepfakes," *Lecture notes in networks and systems*, vol. 1016, pp. 158–166, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-62281-6_11.

[26] M. Brundage *et al.*, "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *arXiv.org*, Feb. 20, 2018. https://arxiv.org/abs/1802.07228

[27] O. O. Olaoye, F. U. Quadri, and O. O. Olaniyi, "Examining the Role of Trade on the Relationship between Environmental Quality and Energy Consumption: Insights from Sub Saharan Africa," *Journal of economics, management and trade*, vol. 30, no. 6, pp. 16–35, Apr. 2024, doi: https://doi.org/10.9734/jemt/2024/v30i61211.

[28] J. Laffier and A. Rehman, "Deepfakes and Harm to Women," *Journal of Digital Life and Learning*, vol. 3, no. 1, pp. 1–21, Jun. 2023, doi: https://doi.org/10.51357/jdll.v3i1.218.

[29] A. Busacca and M. A. Monaca, "Deepfake: Creation, Purpose, Risks," *Studies in systems, decision and control*, vol. 222, pp. 55–68, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-33461-0_6.

[30] A. A. Salami, U. T. I. Igwenagu, C. E. Mesode, O. O. Olaniyi, and O. B. Oladoyinbo, “Beyond Conventional Threat Defense: Implementing Advanced Threat Modeling Techniques, Risk Modeling Frameworks and Contingency Planning in the Healthcare Sector for Enhanced Data Security,” *Journal of Engineering Research and Reports*, vol. 26, no. 5, pp. 304–323, Apr. 2024, doi: https://doi.org/10.9734/jerr/2024/v26i51156.

[31] S. Mengesha, A. Diaz, and K. Dunn, “Protecting Against Sexual Violence Linked to Deepfake Technology | The Regulatory Review,” *www.theregreview.org*, Apr. 13, 2024. https://www.theregreview.org/2024/04/13/protecting-against-sexual-violence-linked-to-deepfake-technology/

[32] S. Briscoe, “U.S Laws Address Deepfakes,” *ASIS*, Jan. 12, 2021. https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/#:~:text=In%20the%20same%20year%2C%20Virginia,against%20deepfakes%20are%20likely%20to (accessed Jun. 28, 2024).

[33] T. O. Oladoyinbo, S. O. Olabanji, O. O. Olaniyi, O. O. Adebiyi, O. J. Okunleye, and A. I. Alao, “Exploring the Challenges of Artificial Intelligence in Data Integrity and its Influence on Social Dynamics,” *Asian Journal of Advanced Research and Reports*, vol. 18, no. 2, pp. 1–23, Jan. 2024, doi: https://doi.org/10.9734/ajarr/2024/v18i2601.

[34] O. R. Ohiro, “WHEN LIES GET REAL: DEEPFAKES AND THE BATTLE FOR REPUTATION IN NIGERIA,” *African Journal Of Law And Human Rights*, vol. 8, no. 1, Jun. 2024, Accessed: Jun. 28, 2024. [Online]. Available: https://www.journals.ezenwaohaetorc.org/index.php/AJLHR/article/view/2864

[35] J. C. Ugonnia, O. O. Olaniyi, F. G. Olaniyi, A. A. Arigbabu, and T. O. Oladoyinbo, “Towards Sustainable IT Infrastructure: Integrating Green Computing with Data Warehouse and Big Data Technologies to Enhance Efficiency and Environmental Responsibility,” *Journal of Engineering Research and Reports*, vol. 26, no. 5, pp. 247–261, Apr. 2024, doi: https://doi.org/10.9734/jerr/2024/v26i51151.

[36] B. van der Sloot and Y. Wagensveld, “Deepfakes: regulatory challenges for the synthetic society,” *Computer Law & Security Review*, vol. 46, p. 105716, Sep. 2022, doi: https://doi.org/10.1016/j.clsr.2022.105716.

[37] A. D. Samuel-Okon and O. O. Abejide, “Bridging the Digital Divide: Exploring the Role of Artificial Intelligence and Automation in Enhancing Connectivity in Developing Nations,” *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 165–177, May 2024, doi: https://doi.org/10.9734/jerr/2024/v26i61170.

[38] A. Flynn, J. Clough, and T. Cooke, “Disrupting and Preventing Deepfake Abuse: Exploring Criminal Law Responses to AI-Facilitated Abuse,” *The Palgrave Handbook of*

Gendered Violence and Technology, pp. 583–603, 2021, doi: https://doi.org/10.1007/978-3-030-83734-1_29.

[39] S. M. Williamson and V. Prybutok, “The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation,” *Information*, vol. 15, no. 6, p. 299, Jun. 2024, doi: https://doi.org/10.3390/info15060299.

[40] P. Yung, N. F. Ma, I.-J. Kim, and D. Yoon, “Speculating on Risks of AI Clones to Selfhood and Relationships: Doppelganger-phobia, Identity Fragmentation, and Living Memories,” *Proceedings of the ACM on human-computer interaction*, vol. 7, no. CSCW1, pp. 1–28, Apr. 2023, doi: https://doi.org/10.1145/3579524.

[41] E. Pashentsev, “The Malicious Use of Deepfakes Against Psychological Security and Political Stability,” *Springer eBooks*, pp. 47–80, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-22552-9_3.

[42] C. S. Adigwe, O. O. Olaniyi, S. O. Olabanji, O. J. Okunleye, N. R. Mayeke, and S. A. Ajayi, “Forecasting the Future: The Interplay of Artificial Intelligence, Innovation, and Competitiveness and its Effect on the Global Economy,” *Asian journal of economics, business and accounting*, vol. 24, no. 4, pp. 126–146, Feb. 2024, doi: https://doi.org/10.9734/ajeba/2024/v24i41269.

[43] S. Karnouskos, “Artificial Intelligence in Digital Media: The Era of Deepfakes,” *IEEE Transactions on Technology and Society*, vol. 1, no. 3, pp. 138–147, 2020, doi: https://doi.org/10.1109/tts.2020.3001312.

[44] Y. Sawada, R. Kanai, and H. Kotani, “Impact of cry wolf effects on social preparedness and the efficiency of flood early warning systems,” *Hydrology and Earth System Sciences*, vol. 26, no. 16, pp. 4265–4278, Aug. 2022, doi: https://doi.org/10.5194/hess-26-4265-2022.

[45] A. T. Arigbabu, O. O. Olaniyi, C. S. Adigwe, O. O. Adebiyi, and S. A. Ajayi, “Data Governance in AI - Enabled Healthcare Systems: A Case of the Project Nightingale,” *Asian Journal of Research in Computer Science*, vol. 17, no. 5, pp. 85–107, Mar. 2024, doi: https://doi.org/10.9734/ajrcos/2024/v17i5441.

[46] Y. A. Marquis, T. O. Oladoyinbo, S. O. Olabanji, O. O. Olaniyi, and S. S. Ajayi, “Proliferation of AI Tools: A Multifaceted Evaluation of User Perceptions and Emerging Trend,” *Asian Journal of Advanced Research and Reports*, vol. 18, no. 1, pp. 30–35, Jan. 2024, doi: https://doi.org/10.9734/ajarr/2024/v18i1596.

[47] J. T. Hancock and J. N. Bailenson, “The Social Impact of Deepfakes,” *Cyberpsychology, Behavior, and Social Networking*, vol. 24, no. 3, pp. 149–152, Mar. 2021, doi: https://doi.org/10.1089/cyber.2021.29208.jth.

[48] Z. Tang, S. Yin, and D. H. Goh, “Understanding Major Topics and Attitudes Toward Deepfakes: An Analysis of News Articles,” *Lecture Notes in Computer Science*, vol. 14056, pp. 337–355, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-48044-7_25.

[49] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas, “Regulating Deep Fakes: Legal and Ethical Considerations,” *papers.ssrn.com*, Dec. 02, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3497144

[50] M. M. Graham, “Deepfakes: Federal and state regulation aims to curb a growing threat,” *Thomson Reuters Institute*, Jun. 26, 2024. https://www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation/ (accessed Jun. 28, 2024).

[51] M. Feeney, “Deepfake Laws Risk Creating More Problems Than They Solve Authored by,” Mar. 2021. Available: https://rtp.fedsoc.org/wp-content/uploads/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf

[52] F. R. Moreno, “Generative AI and deepfakes: a human rights approach to tackling harmful content,” *International review of law computers & technology*, pp. 1–30, Mar. 2024, doi: https://doi.org/10.1080/13600869.2024.2324540.

[53] N. R. Mayeke, A. T. Arigbabu, O. O. Olaniyi, O. J. Okunleye, and C. S. Adigwe, “Evolving Access Control Paradigms: A Comprehensive Multi-Dimensional Analysis of Security Risks and System Assurance in Cyber Engineering. ,” vol. 17, no. 5, pp. 108–124, 2024, doi: https://doi.org/10.9734/ajrcos/2024/v17i5442.

[54] D. Vese, “Governing Fake News: The Regulation of Social Media and the Right to Freedom of Expression in the Era of Emergency,” *European Journal of Risk Regulation*, vol. 13, no. 3, pp. 1–41, Oct. 2021, doi: https://doi.org/10.1017/err.2021.48.

[55] K. Farish, “Do deepfakes pose a golden opportunity? Considering whether English law should adopt California’s publicity right in the age of the deepfake,” *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, Nov. 2019, doi: https://doi.org/10.1093/jiplp/jpz139.

[56] S. O. Olabanji, “AI for Identity and Access Management (IAM) in the Cloud: Exploring the Potential of Artificial Intelligence to Improve User Authentication, Authorization, and Access Control within Cloud-Based Systems,” *Asian Journal of Research in Computer Science*, vol. 17, no. 3, pp. 38–56, 2024, doi: https://doi.org/10.9734/ajrcos/2024/v17i3423.

[57] A. de Ruiter, “The Distinct Wrong of Deepfakes,” *Philosophy & Technology*, vol. 34, pp. 1311–1332, Jun. 2021, doi: https://doi.org/10.1007/s13347-021-00459-2.

[58] S. O. Olabanji, Y. A. Marquis, C. S. Adigwe, A. S. Abidemi, T. O. Oladoyinbo, and O. O. Olaniyi, “AI-Driven Cloud Security: Examining the Impact of User Behavior

Analysis on Threat Detection," *Asian Journal of Research in Computer Science*, vol. 17, no. 3, pp. 57–74, Jan. 2024, doi: https://doi.org/10.9734/ajrcos/2024/v17i3424.

[59] O. O. Olaniyi, O. J. Okunleye, and S. O. Olabanji, "Advancing Data-Driven Decision-Making in Smart Cities through Big Data Analytics: A Comprehensive Review of Existing Literature," *Current Journal of Applied Science and Technology*, vol. 42, no. 25, pp. 10–18, Aug. 2023, doi: https://doi.org/10.9734/cjast/2023/v42i254181.

[60] S. Suratkar, S. Bhiungade, J. Pitale, K. Soni, T. Badgujar, and F. Kazi, "Deep-fake video detection approaches using convolutional – recurrent neural networks," *Journal of Control and Decision*, vol. 10, no. 2, pp. 1–17, Mar. 2022, doi: https://doi.org/10.1080/23307706.2022.2033644.

[61] I. C. Camacho and K. Wang, "A Comprehensive Review of Deep-Learning-Based Methods for Image Forensics," *Journal of Imaging*, vol. 7, no. 4, p. 69, Apr. 2021, doi: https://doi.org/10.3390/jimaging7040069.

[62] A. Kaur, A. N. Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, "Deepfake video detection: challenges and opportunities," *Artificial intelligence review*, vol. 57, no. 6, May 2024, doi: https://doi.org/10.1007/s10462-024-10810-6.

[63] O. O. Olaniyi, O. J. Okunleye, S. O. Olabanji, C. U. Asonze, and S. A. Ajayi, "IoT Security in the Era of Ubiquitous Computing: A Multidisciplinary Approach to Addressing Vulnerabilities and Promoting Resilience," *Asian Journal of Research in Computer Science*, vol. 16, no. 4, pp. 354–371, Dec. 2023, doi: https://doi.org/10.9734/ajrcos/2023/v16i4397.

[64] E. Rohith, B. Padmaja, and V. M. Manikandan, "A Comprehensive Exploration of Advancements and Applications of Digital Watermarking," *Blockchain technologies*, pp. 351–368, Jan. 2024, doi: https://doi.org/10.1007/978-981-97-1249-6_16.

[65] C. Kumar, "Hybrid optimization for secure and robust digital image watermarking with DWT, DCT and SPIHT," *Multimedia tools and applications*, vol. 83, no. 11, pp. 31911–31932, Sep. 2023, doi: https://doi.org/10.1007/s11042-023-16903-8.

[66] O. O. Olaniyi and D. S. Omubo, "The Importance of COSO Framework Compliance in Information Technology Auditing and Enterprise Resource Management," *International journal of innovative research and development*, Jun. 2023, doi: https://doi.org/10.24940/ijird/2023/v12/i5/may23001.

[67] M. Wazid, A. K. Mishra, N. Mohd, and A. K. Das, "A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact," *Cyber Security and Applications*, vol. 2, p. 100040, Feb. 2024, doi: https://doi.org/10.1016/j.csa.2024.100040.

[68] H. Liz-López, M. Keita, A. Taleb-Ahmed, A. Hadid, J. Huertas-Tato, and D. Camacho, “Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges,” *Information Fusion*, vol. 103, p. 102103, Mar. 2024, doi: https://doi.org/10.1016/j.inffus.2023.102103.

[69] O. O. Olaniyi, C. U. Asonze, S. A. Ajayi, S. O. Olabanji, and C. S. Adigwe, “A Regressional Study on the Impact of Organizational Security Culture and Transformational Leadership on Social Engineering Awareness among Bank Employees: The Interplay of Security Education and Behavioral Change,” *Asian Journal of Economics, Business and Accounting*, vol. 23, no. 23, pp. 128–143, Dec. 2023, doi: https://doi.org/10.9734/ajeba/2023/v23i231176.

[70] A. Qureshi and D. Megías Jiménez, “Blockchain-Based Multimedia Content Protection: Review and Open Challenges,” *Applied Sciences*, vol. 11, no. 1, p. 1, Dec. 2020, doi: https://doi.org/10.3390/app11010001.

[71] O. O. Olaniyi, “Ballots and Padlocks: Building Digital Trust and Security in Democracy through Information Governance Strategies and Blockchain Technologies,” *Asian Journal of Research in Computer Science*, vol. 17, no. 5, pp. 172–189, Mar. 2024, doi: https://doi.org/10.9734/ajrcos/2024/v17i5447.

[72] O. O. Olaniyi, S. O. Olabanji, and O. J. Okunleye, “Exploring the Landscape of Decentralized Autonomous Organizations: A Comprehensive Review of Blockchain Initiatives,” *Journal of Scientific Research and Reports*, vol. 29, no. 9, pp. 73–81, Sep. 2023, doi: https://doi.org/10.9734/jsrr/2023/v29i91786.

[73] T. Ciarli, M. Kenney, S. Massini, and L. Piscitello, “Digital technologies, innovation, and skills: Emerging Trajectories and Challenges,” *Research Policy*, vol. 50, no. 6, p. 104289, Jun. 2021.

[74] T. J. Sturgeon, “Upgrading strategies for the digital economy,” *Global Strategy Journal*, vol. 11, no. 1, Dec. 2019.

[75] D. Almeida, K. Shmarko, and E. Lomas, “The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks,” *AI and Ethics*, vol. 2, no. 3, Jul. 2021, Available: https://link.springer.com/article/10.1007/s43681-021-00077-w

[76] A. Mohanarathinam, S. Kamalraj, G. K. D. Prasanna Venkatesan, R. V. Ravi, and C. S. Manikandababu, “Digital watermarking techniques for image security: a review,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 8, pp. 3221–3229, Sep. 2019, doi: https://doi.org/10.1007/s12652-019-01500-1.

[77] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, and Y. Djenouri, “Deepfakes: current and future trends,” *Artificial Intelligence Review*, vol. 57, no. 3, Feb. 2024, doi: https://doi.org/10.1007/s10462-023-10679-x.

[78] G. Bueermann and N. Perucica, “How can we combat the worrying rise in deepfake content?,” *World Economic Forum*, May 19, 2023. https://www.weforum.org/agenda/2023/05/how-can-we-combat-the-worrying-rise-in-deepfake-content/

[79] P. Kadian, S. M. Arora, and N. Arora, “Robust Digital Watermarking Techniques for Copyright Protection of Digital Data: A Survey,” *Wireless Personal Communications*, vol. 118, Feb. 2021, doi: https://doi.org/10.1007/s11277-021-08177-w.

[80] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, “A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods,” *Electronics*, vol. 13, no. 1, pp. 95–95, Dec. 2023, doi: https://doi.org/10.3390/electronics13010095.