GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology

Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

A COMPARATIVE REVIEW OF AI-GENERATED IMAGE DETECTION ACROSS SOCIAL MEDIA PLATFORMS

Anjuman Ara[1], Md Sajadul Alam[2], Kamrujjaman[3] Afia Farjana Mifa[4]

[1]Management Information Systems, College of Business, Beaumont, Texas, USA.
https://orcid.org/0009-0002-1704-8388

[2]Computer Science and Engineering, Department of Computer Science, American International University-Bangladesh, Bangladesh
https://orcid.org/0009-0001-8014-2038

[3]Computer Science and Engineering, Department of Computer Science and Engineering, North South University, Bangladesh
https://orcid.org/0009-0004-4696-921X

[4]Computer Science and Engineering, Department of Computer Science, American International University-Bangladesh, Bangladesh.
https://orcid.org/0009-0001-5827-152X

Abstract

The proliferation of images generated by artificial intelligence (AI) has significantly impacted the digital landscape, especially on social media platforms where the distinction between natural and synthetic content is increasingly blurred. This study embarks on a comparative review of the strategies used by major social media platforms—Facebook/Instagram, Twitter, TikTok, and YouTube—to detect AI-generated images. Employing a comprehensive methodology that includes a systematic review of academic literature, analysis of platform policies, and expert interviews, this research assesses the effectiveness of various detection methods, ranging from sophisticated AI tools to user reporting mechanisms. The findings reveal diverse approaches: Facebook and Instagram utilise a blend of AI detection and human moderation; Twitter integrates machine learning algorithms with user reports; TikTok emphasises AI tools within moderation workflows and educational initiatives; and YouTube relies on its Content ID system alongside AI analysis. The study highlights the critical role of effective detection systems in maintaining content authenticity and user trust, underscoring the importance of balancing automated detection with human oversight. The ongoing development and refinement of these technologies, alongside collaborative efforts and evolving regulatory frameworks, are identified as essential for ensuring a trustworthy digital environment. This research contributes to the discourse on digital integrity, offering insights into the complexities of safeguarding social media ecosystems against the challenges posed by AI-generated content.

Keywords: *AI-generated image detection, Social media content authenticity, Deepfake technology, Content moderation strategies, Digital trust and integrity*

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology
Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

Introduction

The digital landscape has been significantly transformed by the rise of artificial intelligence (AI)-)-generated images have become increasingly prevalent in online media (Tariq et al., 2020). Social media platforms are at the forefront of this transformation, serving as arenas where the authenticity of visual content is constantly tested against the proliferation of synthetic media (Bharati et al., 2016; Hsu et al., 2020). According to Karras et al. (2017), the sophistication of AI techniques, especially those employing Generative Adversarial Networks (GANs), has advanced to a point where the line between genuine and fabricated imagery is often indiscernible. This obscurity poses a substantial challenge to average users and the algorithms designed to detect such content (Zampoglou et al., 2016). As these platforms grapple with the implications of this technological evolution, the need for robust detection mechanisms becomes increasingly critical (Hashmi et al., 2013; Shrivastava et al., 2017). The integrity of these social media ecosystems and the trust users place in them hinges on the ability to discern and authenticate the origins of the content they engage with daily (Shullani et al., 2017).

In response to this emergent challenge, social media giants have begun deploying various strategies to detect AI-generated images and mitigate their potential impacts (Shu et al., 2017). This article examines and contrasts the detection methods currently in use across various platforms. By surveying the landscape of existing technologies, from cutting-edge algorithmic solutions to community-based reporting systems, this analysis aims to shed light on the current state of AI-generated image detection (Parikh & Atrey, 2018; Shu et al., 2017; Y. Wang et al., 2018). The detection of AI-generated content has become a dynamic field, with continuous developments driven by the escalating sophistication of image-generation technologies (McCloskey & Albright, 2019). Platforms are in an ongoing technological arms race, striving to outpace the capabilities of AI generators with more advanced and precise detection algorithms (Singh & Sharma, 2021). This back-and-forth has significant implications for the future of digital content curation and the role of AI in shaping the trustworthiness of shared media (Salim et al., 2022; Sharma et al., 2022).

The importance of understanding and improving AI-generated image detection extends beyond the technical realm; it is a matter that affects the very foundation of how information is perceived and trusted online. With the vast number of users relying on social media for news, personal interactions, and business, the authenticity of visual content has never been more critical. Reliable detection methods are essential to counter the dissemination of deepfakes and other forms of synthetic media that can erode the credibility of online platforms (Elaskily et al., 2021). As AI technology continues to evolve, so must the strategies to combat its misuse. This includes the development of more advanced detection algorithms and the fostering of greater awareness among users about the nature of AI-generated content (Parikh & Atrey, 2018). Ensuring the veracity of shared information in the digital age is a complex, multifaceted endeavour requiring cooperation between technology developers, platform operators, and users (Gaikwad & Hoeber, 2019; Singh & Sharma, 2021). Within this context, this article conducts its comparative analysis, aiming to contribute to the ongoing discussion about maintaining content authenticity in an era increasingly defined by AI.

Background

Artificial intelligence has introduced a groundbreaking method for creating images through algorithms known as Generative Adversarial Networks (GANs)(Salimans et al., 2016). These sophisticated networks consist of a generator that creates images and a discriminator that evaluates their authenticity. Together, they engage in a continuous feedback loop, allowing the generation of images that closely mirror the complexity and detail of real-world visuals (Karras et al., 2017; T.-C. Wang et al., 2018). The use of GANs extends across various domains, including but not limited to

Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

enhancing visual content for films and crafting intricate and lifelike textures for video games (McCloskey & Albright, 2019). Such technology not only showcases the creative potential of AI but also exemplifies the blurring of lines between artificial creations and actual photographic content. The fidelity of these images has progressed to such an extent that it often requires more than a cursory glance to differentiate between an AI-generated image and a genuine photograph, a testament to the rapid advancements in AI capabilities (Guarnera et al., 2020; Hsu et al., 2020). As GANs continue to evolve and become more accessible, their applications are likely to expand further, permeating more sectors and presenting new challenges and opportunities within digital content creation.

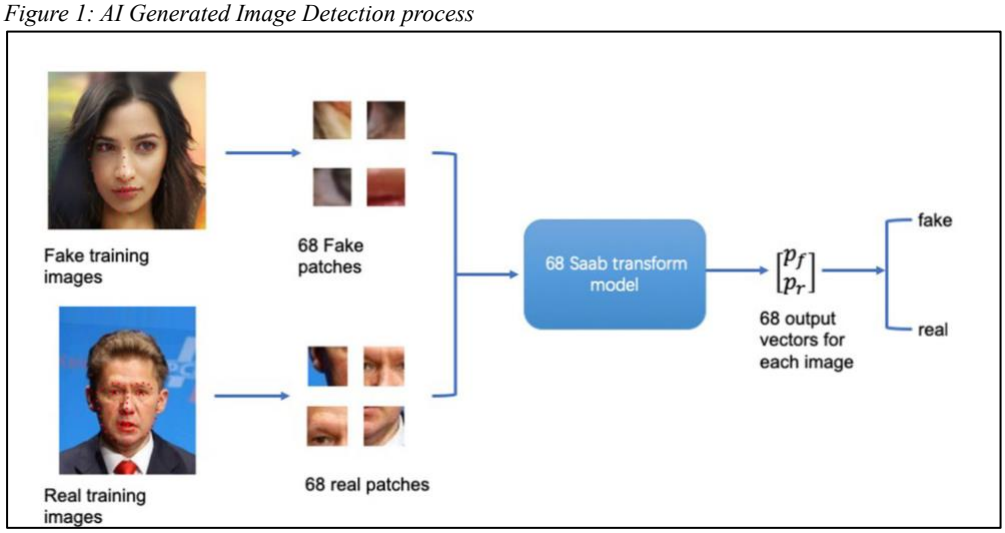
The verisimilitude of AI-generated images has introduced considerable complexity to the issue of content authenticity on social media platforms. This evolving technology enables the creation of deepfakes, which are alarmingly convincing counterfeit videos or images (Jeon et al., 2020; Jeon et al., 2020). Such deepfakes can potentially impersonate public figures, fabricate scenarios, or sway public sentiment, thus posing a severe threat to the integrity of information circulated online. The implications of this technological misuse are profound, as they undermine users' trust in digital content. As a result, the distinction between authentic and AI-generated content has become an increasingly common challenge for social media users, leading to a pressing need for reliable verification methods. The uncertainty surrounding the authenticity of online content necessitates a vigilant approach to media consumption and a heightened awareness of the capabilities of AI in image generation (Neves et al., 2020; Shrivastava et al., 2017). With the potential of such technologies to influence social, political, and personal realms, the detection and management of AI-generated content has become a critical concern for maintaining the credibility of digital platforms. Building on the work of Zhu (2019) and the field defined by pioneers such as Li et al. (2007), the study advances the domain of AI-generated image detection by proposing an interpretable and less computationally intensive methodology. Zhu's approach diverges from the conventional reliance on deep learning and CNNs by employing a two-layer Saab transform model to analyse image patches centred on facial landmarks. This technique mitigates the opacity that typically characterises CNN-based methods, offering a window into the decision-making process that underpins the classification of images as real or fake (Zhu, 2019). The process begins with extracting 68 facial landmarks from real and fake training images. These landmarks are focal points for extracting 32x32 pixel patches, which are then analysed by the Saab transform model to determine the image's authenticity (Zhu, 2019). The output is a series of 68 2x1 vectors representing the probability of each patch being real or fake. These vectors are subsequently fed into a Support Vector Machine (SVM) classifier, synthesising the data to render a final verdict on the image's authenticity. This method's transparency not only provides clear insights into its analytic process but also aligns with the growing demand for understandable AI within the realm of digital content management.

Figure 1 illustrates this innovative process, showcasing the flow from the initial image input through the facial landmark extraction, patch analysis, and culminating in the SVM classification. By training separate models for each landmark, Zhu's approach capitalizes on the granularity of facial features, a detail often lost in broader CNN analyses. This methodology's potential to achieve accuracy on par with more complex systems, without sacrificing interpretability, is particularly valuable in the context of social media, where the provenance of an image can have significant implications for public perception and discourse. Incorporating Zhu's (2019) model into the detection frameworks of social media platforms could vastly enhance their ability to discern genuine content from AI-generated forgeries. As these platforms continue to grapple with the deluge of synthetic media, methodologies like Zhu's provide a beacon for developing detection systems that are both effective and user-friendly. This advancement in technology paves the way for a future where digital content

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology
Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

can be consumed with confidence, and the authenticity of media is preserved in the face of increasingly sophisticated AI-generated images (See Figure 1).

Figure 1: AI Generated Image Detection process



Source: Zhu (2019)

To tackle the complexities presented by AI-generated content, the tech industry has rallied to develop an array of detection technologies (Sheng et al., 2018). At the forefront are deep learning models, renowned for their proficiency in pattern recognition, which have been fine-tuned to identify the subtlest of anomalies characteristic of fabricated images. These sophisticated models scrutinize for peculiarities in elements such as texture, lighting, and edge delineation—attributes that typically betray an image's synthetic origins and may elude the human eye's detection (Tahaoglu et al., 2022; Tanaka et al., 2021; Tyagi & Yadav, 2022). In addition to visual analysis, some strategies delve into the digital DNA of images by examining their metadata or tracing the origins of their distribution (Wu et al., 2022). Others have homed in on the distinctive digital 'fingerprints' that generative processes imprint on images. Yet, despite the progress made, the battle to discern AI-generated images is in constant flux. The creation and detection techniques are locked in a perpetual cycle of one-upmanship: as generators become more adept at mimicking reality, detectors must correspondingly escalate their capabilities to unmask these ever-more-convincing digital facades (Zhang et al., 2023). This ongoing contest not only fuels technological advancement but also raises the stakes in ensuring the integrity of shared media content.

The relentless advancement in AI technology propels the task of preserving content authenticity into an ever-evolving challenge (Tyagi & Yadav, 2022). Social media platforms are finding themselves at the crux of this issue, tasked with the responsibility of safeguarding user trust and the integrity of the content shared within their domains. To achieve this, platforms are compelled to employ state-of-the-art detection algorithms that can keep up with the rapid pace of AI-generated image creation. However, technology alone may not suffice (Niu et al., 2021; Zhou et al., 2018). There is a growing recognition of the need for transparency in the mechanisms of detection and moderation, as well as an emphasis on user education. Users equipped with knowledge about the nature of AI-generated content and the means to identify it can serve as an invaluable line of defence against the spread of inauthentic content ((Sun et al., 2022; Tahaoglu et al., 2022). As such, the responsibility of platforms extends beyond the technical to encompass the fostering of digital literacy, ensuring that users are well-informed participants in the digital ecosystem. This dual approach, combining technological solutions with informed user engagement, is becoming increasingly essential to navigate the

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology
Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

intricacies of digital media and to maintain the credibility that is fundamental to the social media experience.

Methodology

The methodology underpinning this study involved a comprehensive and systematic review designed to evaluate the detection methods for AI-generated images across several prominent social media platforms. The review commenced with a meticulous search of academic databases and platform-specific documentation to gather relevant literature and policy statements concerning the deployment of AI detection tools and moderation strategies. A selection criterion was rigorously applied to ensure the inclusion of studies that specifically addressed AI-generated content detection and user engagement in content moderation processes. Data collection was augmented through expert interviews to provide qualitative insights into the operational challenges and effectiveness of each platform's detection methods. Following data synthesis, a comparative analysis was conducted, assessing the platforms' reliance on AI tools, the role of human moderators, user involvement in reporting suspicious content, and the impact of these combined efforts on maintaining content integrity. The evaluation culminated in the construction of a summarizing table (Table 1) that provides an at-a-glance comparison of the various strategies, their effectiveness, and the implications for users and content creators. This methodical approach facilitated a nuanced understanding of each platform's tactics in managing the ever-evolving challenge of AI-generated synthetic media.

Findings

Findings from the systematic review of AI-generated image detection across various social media platforms reveal distinct approaches employed by each platform to tackle the challenges posed by synthetic media.

Facebook/Instagram

On the platforms of Facebook and Instagram, a hybrid approach has been adopted to ensure the authenticity of content, where sophisticated AI tools work in tandem with human moderation teams. The AI employed on these platforms is equipped with advanced pattern recognition and anomaly detection capabilities, designed to meticulously scan and identify potential AI-generated images that could breach the platforms' integrity. When such images are detected, they are flagged for further inspection, which triggers the intervention of human moderators. This second layer of defense is where the nuances of content are evaluated, allowing for the discernment that AI alone may not possess. Human moderators examine the flagged content with a critical eye, considering context, cultural references, and subtleties that the AI may overlook. This balanced methodology is crucial to the process, as it seeks to harness the speed and scalability of AI while ensuring that the accuracy and fairness of content moderation are upheld by human insight. This symbiotic relationship between AI and human judgment is essential in maintaining a seamless user experience, preserving trust, and safeguarding the platforms from the infiltration of deceptive content. It epitomizes the platforms' commitment to maintaining a genuine space for user interaction, one where efficiency in content moderation does not compromise the thoroughness required in the era of sophisticated digital content creation.

Twitter

Twitter has harnessed the power of machine learning algorithms to create an automated defense against the proliferation of AI-generated content on its platform. These algorithms, steeped in extensive training from expansive datasets, are fine-tuned to detect the faint but telltale signs that characterize synthetic media. The nuanced detection process involves sifting through the digital noise to pinpoint irregularities that may suggest an image's inauthenticity. Beyond the realm of

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology
Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

algorithmic oversight, Twitter also taps into the collective vigilance of its user base through a user-reporting mechanism. This feature democratizes the process of content moderation, allowing users to act as sentinels who can flag content that raises suspicion. Such a participatory approach not only extends the reach of Twitter's content policing but also fosters a sense of community responsibility. Together, the machine precision of algorithmic flagging and the human intuition of user reports create a layered shield against the subtle infiltration of AI-generated images. This strategy reflects Twitter's acknowledgment of the complexity of content verification in the digital age and its commitment to leveraging both technological and human resources to maintain the authenticity of shared content.

TikTok

TikTok's approach to safeguarding its platform against the encroachment of AI-generated content involves a seamless integration of AI detection tools within its content moderation workflows. These tools are not merely passive filters but active scanners that meticulously analyze each piece of uploaded content, probing for indicators of artificial genesis. They scrutinize visual cues that might betray the handiwork of generative algorithms and examine metadata that could reveal traces of digital manipulation. Beyond the deployment of these technological sentinels, TikTok recognizes the power of an informed community. To this end, the platform has embarked on a mission to educate its users, developing initiatives designed to illuminate the hallmarks of AI-generated content. Through tutorials, guidelines, and interactive features, TikTok empowers its users to become active participants in the moderation process. By fostering an environment where users are not only alert to the potential of artificial content but also equipped to recognize its subtleties, TikTok enhances its defenses against the dissemination of synthetic media. This educational strategy is pivotal, as it cultivates a user base that is vigilant, knowledgeable, and engaged in the collective endeavor to maintain the authenticity of the content ecosystem.

YouTube

YouTube’s Content ID system stands as a testament to the platform's commitment to intellectual property rights and the authenticity of its content, harnessing advanced AI to scrutinize videos at the point of upload. This AI is not simply a tool for detecting copyright infringement; it has evolved to become a sentinel against AI-generated videos, distinguishing between genuine creations and those that may have been artificially fabricated. By analyzing the myriad of visual and audio signals within each video, the AI delves into patterns that human eyes or ears might miss, searching for discrepancies indicative of synthetic production. While this technological prowess is commendable, it brings with it a significant responsibility towards content creators whose livelihoods often hinge on the platform. The AI’s discernment is crucial, as inaccuracies can lead to false positives, resulting in the demonetization or unwarranted removal of videos, disrupting the creators' revenue streams and potentially stifling creative expression. The precision of YouTube’s AI system is, therefore, of paramount importance, as it navigates the delicate balance between safeguarding against synthetic media and upholding the rights and revenues of genuine content creators. This balance is a cornerstone of YouTube's platform, ensuring that while it remains a bastion against digital counterfeits, it also continues to be a space where creators can flourish without fear of undue penalization.

Comparative Analysis

The comparative analysis of AI-generated image detection strategies across major social media platforms demonstrates a spectrum of effectiveness influenced by the complexity and training quality of AI algorithms. On platforms like Facebook and Instagram, pattern recognition and anomaly detection are paramount, with human moderation providing essential oversight, particularly for content flagged by AI, thus ensuring a balance between technological efficiency and the nuances of

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology
Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

human judgment. Twitter's model leans heavily on machine learning algorithms for initial flagging, augmented by a user-reporting system that, while empowering users to partake in content moderation, introduces the potential for bias and error. TikTok's approach is notable for embedding AI detection tools directly within its content moderation workflow, complemented by educational initiatives aimed at enhancing user discernment regarding AI-generated content. YouTube's Content ID system represents a dual focus on copyright protection and the detection of synthetic media, relying on AI for initial analysis and human moderators for final adjudication, a process that directly impacts content monetization and underscores the need for accuracy in the AI's decision-making process. Across these platforms, the interplay between automated AI tools, human moderation, and user engagement forms a multi-layered defense against the incursion of AI-generated content, each with its own strengths and challenges in the face of scalable content management and the continuous evolution of generative AI techniques (See Table 1).

Table 1: overview of the different strategies employed by each platform.

Platform	AI Detection	Human Moderation	User Involvement	Impact on Users
Facebook/Instagram	Pattern recognition and anomaly detection	Critical for reviewing AI-flagged content	Limited to reporting suspicious content	Balances efficiency with oversight
Twitter	Machine learning algorithms for flagging	Secondary to AI, with user-reporting as backup	User-reporting mechanisms for flagging	Empowers users, but susceptible to bias
TikTok	AI detection tools in content moderation workflows	Part of the workflow, but with emphasis on AI tools	Educational initiatives to inform users	Raises awareness, fosters discerning viewers
YouTube	Content ID system and AI for pattern analysis	Essential for appeals and complex decisions	Dependent on users for flagging and feedback	Can affect monetization due to false positives

Discussion

The integrity of content on social media platforms is foundational to user trust, a principle that is becoming increasingly significant as AI-generated images become more prevalent (Matern et al., 2019; Selvaraju et al., 2017). Trust in the authenticity of content is not only crucial for user engagement but also for the overall vitality of social media ecosystems (Durall et al., 2019; Matern et al., 2019; Selvaraju et al., 2017). Effective detection methods serve as a bulwark against the spread of (Jalab et al., 2022; Rana et al., 2022; Singh et al., 2020), which carries the potential to skew public discourse and influence socio-political dynamics (He et al., 2018). Platforms that demonstrate the ability to manage the authenticity of their content transparently and reliably are better positioned to sustain and expand their user base. Conversely, the suspicion of manipulated content can erode user trust and diminish platform credibility, potentially leading to reduced engagement (Nguyen et al., 2022). Therefore, the adoption and implementation of advanced detection technologies are imperative, transcending beyond a mere technical pursuit to becoming a strategic component that underpins user retention and the preservation of platform integrity.

However, the introduction of sophisticated detection technologies to identify AI-generated images is fraught with challenges, particularly the risks of false positives and negatives. False positives can result in unwarranted censorship or content removal, adversely affecting creators' visibility and curtailing their freedom of expression—a situation that could lead to revenue loss, diminished audience reach, and reputational damage (Warif et al., 2015). On the flip side, false negatives could permit the circulation of deepfakes or misinformation, thus compromising the reliability of content disseminated on social media platforms. The accuracy of these detection methods is not just a technological imperative but an ethical one, as it significantly impacts the landscape of digital information (Fan et al., 2017). Ensuring precision in detection is therefore critical, as the consequences of failure in this area are far-reaching, affecting not only individual users but the broader societal trust in digital platforms.

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology
Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

The role of human oversight in the content verification process is indispensable. While AI is proficient in managing and analyzing large volumes of content, it inherently lacks the capability for contextual interpretation and cultural nuance (Boulkenafet et al., 2015; Holmes et al., 2016; Martín-Rodríguez et al., 2023). Human moderators are vital for filling these gaps, bringing an understanding of context, sarcasm, and cultural subtleties to the table—nuances that AI systems may misinterpret or overlook entirely. The importance of human intervention in content moderation cannot be overstated, as it ensures that legitimate content is not incorrectly penalized and that the complexities of human expression are accurately considered (Ferrara et al., 2012; Tariq et al., 2020; van den Oord et al., 2016). The interplay between automated systems and human moderation is delicate, with platforms striving to find the right balance between efficiency and precision. The ultimate goal for social media platforms is to foster a collaborative environment where AI and human moderators operate synergistically, combining the scalability of automation with the discernment of human judgment to establish a digital space that is both secure and conducive to genuine expression (Gaikwad & Hoeber, 2019; Sharma et al., 2022; Shu et al., 2017).

Future Directions

The increasing sophistication of AI-generated content heralds a future where the ability to distinguish authentic from synthetic media is paramount, driving an escalating contest between the technologies used to create and detect such content. Social media's landscape will be inexorably shaped by the progress of AI detection technologies, necessitating persistent innovation and research. This evolution demands an interdisciplinary approach that harnesses cutting-edge machine learning, digital forensics, and cognitive psychology to not only keep pace with but also preempt the next generation of generative AI. Collaborative synergies are vital, with social media entities, technology firms, and academic circles merging their respective data-rich resources, practical insights, and theoretical acumen to pioneer refined detection methodologies. Such partnerships are poised to catalyze significant strides in creating detection models that more closely mirror the intricacies of human-created content. Concurrently, the emergence of rigorous regulatory frameworks and ethical standards plays a crucial role in this landscape, as they must evolve to encapsulate the intricacies of AI's role in our daily digital interactions. These frameworks and guidelines are tasked with the dual responsibility of mitigating the malicious applications of AI while safeguarding individual liberties and ensuring equitable practices within AI systems. The collective journey towards dependable detection of AI-generated content is complex and layered, intertwining continual technological strides, collaborative dynamism, and stringent oversight. The delicate balance struck in this tripartite interaction will ultimately determine the capacity of social media platforms to curate spaces where authenticity prevails, and user trust in the content they engage with remains unshaken.

Conclusion

The comparative review of AI-generated image detection methods across social media giants like Facebook, Instagram, Twitter, TikTok, and YouTube has unveiled a varied array of strategies each platform employs to mitigate the spread of synthetic content. These platforms leverage AI detection technology for its capacity to efficiently process and analyze vast quantities of data, utilizing pattern recognition and anomaly detection algorithms to identify potential deepfakes. However, the reliance on AI is balanced with human moderators, who are crucial for their ability to provide context-sensitive analysis—interpreting nuances and cultural references that automated systems might miss. Additionally, user participation through reporting mechanisms plays a supportive role, though it carries the inherent risk of introducing bias. The efficacy of these detection systems is pivotal for maintaining content authenticity and securing user trust, as the line between AI-generated and genuine content becomes increasingly blurred. Social media platforms, serving as forums for public

Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

and private expression, must persistently innovate their detection technologies to protect their integrity. The synthesis of automated systems with human oversight is essential for content moderation that is both efficient and discerning. Future efforts must also include collaborative initiatives and the development of regulatory and ethical guidelines to ensure the platforms can provide a reliable and safe digital environment. The continued evolution of AI content and the dedication to ethical innovation are critical for ensuring the online experience remains authentic and secure for users worldwide.

References

1. Bharati, A., Singh, R., Vatsa, M., & Bowyer, K. W. (2016). Detecting Facial Retouching Using Supervised Deep Learning. *IEEE Transactions on Information Forensics and Security*, *11*(9), 1903-1913. https://doi.org/10.1109/tifs.2016.2561898
2. Boulkenafet, Z., Komulainen, J., & Hadid, A. (2015). ICIP - Face anti-spoofing based on color texture analysis. *2015 IEEE International Conference on Image Processing (ICIP)*, *NA*(NA), 2636-2640. https://doi.org/10.1109/icip.2015.7351280
3. Durall, R., Keuper, M., Pfreundt, F.-J., & Keuper, J. (2019). Unmasking DeepFakes with simple Features. *arXiv: Learning*, *NA*(NA), NA-NA. https://doi.org/NA
4. Elaskily, M. A., Alkinani, M. H., Sedik, A., & Dessouky, M. M. (2021). Deep learning based algorithm (ConvLSTM) for Copy Move Forgery Detection. *Journal of Intelligent & Fuzzy Systems*, *40*(3), 4385-4405. https://doi.org/10.3233/jifs-201192
5. Fan, S., Ng, T.-T., Koenig, B. L., Herberg, J. S., Jiang, M., Shen, Z., & Zhao, Q. (2017). Image Visual Realism: From Human Perception to Machine Computation. *IEEE transactions on pattern analysis and machine intelligence*, *40*(9), 2180-2193. https://doi.org/10.1109/tpami.2017.2747150
6. Ferrara, P., Bianchi, T., De Rosa, A., & Piva, A. (2012). Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts. *IEEE Transactions on Information Forensics and Security*, *7*(5), 1566-1577. https://doi.org/10.1109/tifs.2012.2202227
7. Gaikwad, M., & Hoeber, O. (2019). CHIIR - An Interactive Image Retrieval Approach to Searching for Images on Social Media. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, *NA*(NA), 173-181. https://doi.org/10.1145/3295750.3298930
8. Guarnera, L., Giudice, O., & Battiato, S. (2020). Fighting Deepfake by Exposing the Convolutional Traces on Images. *IEEE Access*, *8*(NA), 165085-165098. https://doi.org/10.1109/access.2020.3023037
9. Hashmi, M. F., Hambarde, A. R., & Keskar, A. G. (2013). ISDA - Copy move forgery detection using DWT and SIFT features. *2013 13th International Conference on Intellient Systems Design and Applications*, *NA*(NA), 188-193. https://doi.org/10.1109/isda.2013.6920733
10. He, P., Jiang, X., Sun, T., & Li, H. (2018). Computer Graphics Identification Combining Convolutional and Recurrent Neural Networks. *IEEE Signal Processing Letters*, *25*(9), 1369-1373. https://doi.org/10.1109/lsp.2018.2855566
11. Holmes, O., Banks, M. S., & Farid, H. (2016). Assessing and Improving the Identification of Computer-Generated Portraits. *ACM Transactions on Applied Perception*, *13*(2), 7-12. https://doi.org/10.1145/2871714
12. Hsu, C.-C., Zhuang, Y. X., & Lee, C.-Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Applied Sciences*, *10*(1), 370-NA. https://doi.org/10.3390/app10010370
13. Jalab, H. A., Alqarni, M. A., Ibrahim, R. W., & Ali Almazroi, A. (2022). A novel pixel's fractional mean-based image enhancement algorithm for better image splicing detection. *Journal of King Saud University - Science*, *34*(2), 101805-101805. https://doi.org/10.1016/j.jksus.2021.101805
14. Jeon, H., Bang, Y., & Woo, S. S. (2020). SEC - FDFtNet: Facing Off Fake Images using Fake Detection Fine-tuning Network. In (Vol. 580, pp. 416-430). https://doi.org/10.1007/978-3-030-58201-2_28

Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

15. Jeon, H., Bang, Y. O., Kim, J. S., & Woo, S. S. (2020). T-GD: Transferable GAN-generated Images Detection Framework. *arXiv: Computer Vision and Pattern Recognition*, *NA*(NA), NA-NA. https://doi.org/NA
16. Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *arXiv: Neural and Evolutionary Computing*, *NA*(NA), NA-NA. https://doi.org/NA
17. Li, G., Wu, Q., Tu, D., & Sun, S. (2007). ICME - A Sorted Neighborhood Approach for Detecting Duplicated Regions in Image Forgeries Based on DWT and SVD. *Multimedia and Expo, 2007 IEEE International Conference on*, *NA*(NA), 1750-1753. https://doi.org/10.1109/icme.2007.4285009
18. Martín-Rodríguez, F., Isasi-de-Vicente, F., & Fernández-Barciela, M. (2023). A Stress Test for Robustness of Photo Response Nonuniformity (Camera Sensor Fingerprint) Identification on Smartphones. *Sensors (Basel, Switzerland)*, *23*(7), 3462-3462. https://doi.org/10.3390/s23073462
19. Matern, F., Riess, C., & Stamminger, M. (2019). *WACV Workshops - Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations* (Vol. NA). https://doi.org/10.1109/wacvw.2019.00020
20. McCloskey, S., & Albright, M. (2019). ICIP - Detecting GAN-Generated Imagery Using Saturation Cues. *2019 IEEE International Conference on Image Processing (ICIP)*, *NA*(NA), 4584-4588. https://doi.org/10.1109/icip.2019.8803661
21. Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., & Fierrez, J. (2020). GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 1038-1048. https://doi.org/10.1109/jstsp.2020.3007250
22. Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*(NA), 103525-103525. https://doi.org/10.1016/j.cviu.2022.103525
23. Niu, P.-p., Wang, C.-p., Chen, W., Yang, H.-Y., & Wang, X.-y. (2021). Fast and effective Keypoint-based image copy-move forgery detection using complex-valued moment invariants. *Journal of Visual Communication and Image Representation*, *77*(NA), 103068-NA. https://doi.org/10.1016/j.jvcir.2021.103068
24. Parikh, S. B., & Atrey, P. K. (2018). *MIPR - Media-Rich Fake News Detection: A Survey* (Vol. NA). https://doi.org/10.1109/mipr.2018.00093
25. Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*, *10*(NA), 25494-25513. https://doi.org/10.1109/access.2022.3154404
26. Salim, M. Z., Abboud, A. J., & Yildirim, R. (2022). A Visual Cryptography-Based Watermarking Approach for the Detection and Localization of Image Forgery. *Electronics*, *11*(1), 136-136. https://doi.org/10.3390/electronics11010136
27. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved Techniques for Training GANs. *arXiv: Learning*, *NA*(NA), NA-NA. https://doi.org/NA
28. Schetinger, V., Oliveira, M. M., da Silva, R., & Carvalho, T. (2017). Humans Are Easily Fooled by Digital Images. *Computers & Graphics*, *68*(NA), 142-151. https://doi.org/10.1016/j.cag.2017.08.010
29. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). ICCV - Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, *NA*(NA), 618-626. https://doi.org/10.1109/iccv.2017.74
30. Shamim, M. M. I. (2022). Cloud Computing and AI in Analysis of Worksite. *Nexus: Journal of Advances Studies of Engineering Science*, *1*(3), 1-9.
31. Sharma, D. K., Singh, B., Agarwal, S., Kim, H., & Sharma, R. (2022). Sarcasm Detection over Social Media Platforms Using Hybrid Auto-Encoder-Based Model. *Electronics*, *11*(18), 2844-2844. https://doi.org/10.3390/electronics11182844

Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

32. Sheng, H., Shen, X., Yingda, L., Shi, Z., & Ma, S. (2018). Image splicing detection based on Markov features in discrete octonion cosine transform domain. *IET Image Processing*, *12*(10), 1815-1823. https://doi.org/10.1049/iet-ipr.2017.1131
33. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J. M., Wang, W., & Webb, R. (2017). CVPR - Learning from Simulated and Unsupervised Images through Adversarial Training. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *NA*(NA), 2242-2251. https://doi.org/10.1109/cvpr.2017.241
34. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22-36. https://doi.org/10.1145/3137597.3137600
35. Shullani, D., Fontani, M., Iuliani, M., Al Shaya, O., & Piva, A. (2017). VISION: a video and image dataset for source identification. *EURASIP Journal on Information Security*, *2017*(1), 1-16. https://doi.org/10.1186/s13635-017-0067-2
36. Singh, B., & Sharma, D. K. (2021). Predicting image credibility in fake news over social media using multi-modal approach. *Neural computing & applications*, *34*(24), 1-15. https://doi.org/10.1007/s00521-021-06086-4
37. Singh, V. K., Ghosh, I., & Sonagara, D. (2020). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, *72*(1), 3-17. https://doi.org/10.1002/asi.24359
38. Sun, Y., Ni, R., & Zhao, Y. (2022). ET: Edge-Enhanced Transformer for Image Splicing Detection. *IEEE Signal Processing Letters*, *29*(NA), 1232-1236. https://doi.org/10.1109/lsp.2022.3172617
39. Tahaoglu, G., Ulutas, G., Ustubioglu, B., Ulutas, M., & Nabiyev, V. V. (2022). Ciratefi based copy move forgery detection on digital images. *Multimedia Tools and Applications*, *81*(16), 22867-22902. https://doi.org/10.1007/s11042-021-11503-w
40. Tanaka, M., Shiota, S., & Kiya, H. (2021). A Detection Method of Operated Fake-Images Using Robust Hashing. *Journal of imaging*, *7*(8), 134-NA. https://doi.org/10.3390/jimaging7080134
41. Tariq, S., Lee, S., & Woo, S. S. (2020). A Convolutional LSTM based Residual Network for Deepfake Video Detection. *arXiv: Computer Vision and Pattern Recognition*, *NA*(NA), NA-NA. https://doi.org/NA
42. Tyagi, S., & Yadav, D. (2022). MiniNet: a concise CNN for image forgery detection. *Evolving Systems*, *14*(3), 545-556. https://doi.org/10.1007/s12530-022-09446-0
43. van den Oord, A., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., & Kavukcuoglu, K. (2016). Conditional Image Generation with PixelCNN Decoders. *arXiv: Computer Vision and Pattern Recognition*, *NA*(NA), NA-NA. https://doi.org/NA
44. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). CVPR - High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *NA*(NA), 8798-8807. https://doi.org/10.1109/cvpr.2018.00917
45. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). KDD - EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, *NA*(NA), 849-857. https://doi.org/10.1145/3219819.3219903
46. Warif, N. B. A., Idris, M. Y. I., Wahab, A. W. A., & Salleh, R. (2015). An evaluation of Error Level Analysis in image forensics. *2015 5th IEEE International Conference on System Engineering and Technology (ICSET)*, *NA*(NA), 23-28. https://doi.org/10.1109/icsengt.2015.7412439
47. Wu, H., Zhou, J., Tian, J., Liu, J., & Qiao, Y. (2022). Robust Image Forgery Detection Against Transmission Over Online Social Networks. *IEEE Transactions on Information Forensics and Security*, *17*(NA), 443-456. https://doi.org/10.1109/tifs.2022.3144878
48. Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. (2016). Large-scale evaluation of splicing localization algorithms for web images. *Multimedia Tools and Applications*, *76*(4), 4801-4834. https://doi.org/10.1007/s11042-016-3795-2

GMJ Global Mainstream Journal of Innovation, Engineering & Emerging Technology

Volume: 03
Issue: 01
ISSN ONLINE: 2834-2739
February 2024
Texas, USA

49. Zhang, Y., Tan, Q., Qi, S., & Xue, M. (2023). PRNU-based Image Forgery Localization with Deep Multi-scale Fusion. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *19*(2), 1-20. https://doi.org/10.1145/3548689
50. Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2018). CVPR - Learning Rich Features for Image Manipulation Detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *NA*(NA), 1053-1061. https://doi.org/10.1109/cvpr.2018.00116
51. Zhu, Y. (2019). *MCL Research on Fake Image Detection*. https://mcl.usc.edu/news/2019/04/08/mcl-research-on-fake-image-detection/