International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Beyond Illusions: Contribution of Artificial Intelligence in Unveiling and Mitigating Deep Fake Impact on Social Networks

Virender Dhiman*

Submitted: 06/02/2024 **Revised**: 14/03/2024 **Accepted**: 20/03/2024

Abstract: This article investigates the effects of deepfake technology on social networks and evaluates AI-based mitigating strategies. Deepfakes, or synthetic media created by AI, present concerns such as misrepresentation and privacy breaches. Deepfake detection, Face Manipulation Detection Networks (FMDNs) and multimodal analysis are the important AI techniques. Real-world implementations demonstrate less deepfake diffusion. Future research will focus on model interpretability, multidisciplinary cooperation, and media literacy in order to effectively mitigate deepfakes. Ethical issues are critical in addressing emerging challenges.

Keywords: *Deepfake, AI-based mitigation, disinformation, privacy, detection algorithms, multimodal, multidisciplinary cooperation, media literacy*

I. INTRODUCTION

The widespread use of deepfake technology presents notable obstacles to the legitimacy and dependability of digital material on social media platforms. The use of synthetic media that manipulates or fabricates a person's resemblance, or "deepfakes," has increased dramatically as a result of advances in machine learning (ML) and artificial intelligence (AI) algorithms. The spread of false information, fake news, and harmful content has increased as a result of this problem, affecting public opinion and jeopardising the reliability of digital communication systems. The rise of deepfakes on social media has raised questions about the accuracy of content and the possibility of mass deceit. The distinction between authentic and manipulated material becomes increasingly hazy as AI-driven algorithms get more complex, which has significant ramifications for online discourse, the development of public opinion, and public confidence in digital media.

Data emphasises how serious the deepfake problem is. A cybersecurity firm that specialises in deepfake detection, called Deeptrace, said that between 2018 and 2019, the quantity of deepfake movies on the internet quadrupled,

reaching over 14,000 films worldwide. Furthermore, a Pew Research Centre poll found that 63% of participants thought that deepfake videos had a big influence on political discourse, underscoring the widespread impact of manipulated media on public opinion and democratic processes. In addition to these numbers, a research conducted by Sensity—a visual threat intelligence firm driven by artificial intelligence—found that, between 2019 and 2021, the number of deepfake videos discovered increased by an astounding 330% across a variety of social media platforms. Additionally, according to the same study, deepfake films have up to 10 times the amount of engagement compared to non-deepfake content, which increases their ability to mislead gullible viewers.

II.LITERATURE REVIEW

Deepfakes, which are artificial intelligence (AI)-generated synthetic media that modify or falsify material, are a developing worry because of the potential effects they may have on social networks. Scholars and researchers have thoroughly investigated a number of deepfake aspects, including as mitigation techniques, regulatory frameworks, detection techniques, and societal ramifications.

The advancement and efficacy of AI-driven deepfake detection methods is a crucial field of study. The importance of machine learning algorithms—in particular, convolutional neural networks (CNNs)—in recognising visual abnormalities suggestive of deepfake manipulation was highlighted by (Ahmed et al., 2022). (Remya Revi et al., 2021) have demonstrated the efficacy of generative adversarial networks (GANs) in identifying minute discrepancies in the motions and facial expressions of deepfake movies.

Researchers have studied the underlying AI principles used by malicious actors in deepfake generating approaches. (Rahman et al., 2022) talked about the

[*]*Independent Researcher, TX, USA, ORCID: 0009-0002-7429-8703*
[*]*Corresponding Author Email: vdhiman2@illinois.edu*

developments in deep learning models that make it possible to create high-fidelity deepfake material, such autoencoders and recurrent neural networks (RNNs). Furthermore, (Kietzmann et al., 2020) emphasised the significance of responsible AI development techniques by examining the ethical conundrums related to deepfake production.

Understanding the effects of deepfakes on social networks in the actual world has also been made possible by case studies and examples. (Montasari, 2024) conducted an analysis on the consequences of deepfake films in political settings, emphasising their capacity to sway public opinion and jeopardise democratic procedures. Furthermore, Kim et al. (2021) investigated the use of deepfakes in the advertising and entertainment sectors, posing questions about consumer authenticity and trust.

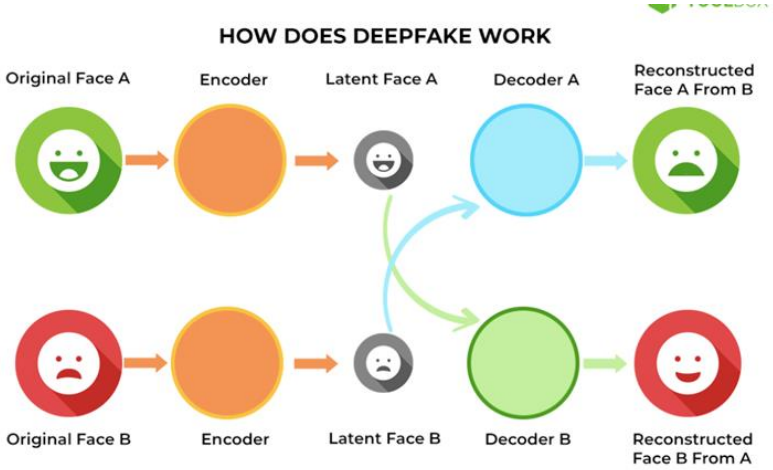

Fig 2.1: How Deep fake works

(https://images.spiceworks.com/wp-content/uploads/2022/05/23151920/How-Does-Deepfake-Work.png)

Artificial Intelligence-based mitigation solutions have attracted a lot of attention in response to the deepfake problem. According to (Ferrara, 2024) proactive detection and removal of deepfake content on social media platforms is made possible by the incorporation of AI algorithms into content moderation systems. Moreover, frameworks for laws and regulations have been put out to address the privacy and ethical issues related to deepfakes (Helmus, 2022).

Research continues to be focused on potential future trends and difficulties in preventing the influence of deepfakes on social networks. (Köbis et al., 2021) projected how AI-powered solutions will develop, highlighting the necessity of ongoing innovation to keep up with advances in deepfake technology. Furthermore, (Mubarak et al., 2023) emphasise that responsible AI practices and the integrity of digital communication channels depend heavily on cooperation between academics, industry, and policymakers.

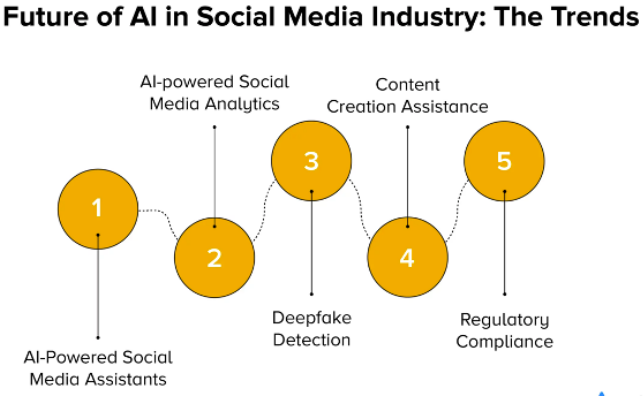

Fig 2.2: Trends of AI in Social Media

(https://appinventiv.com/wp-content/uploads/2023/10/How-Artificial-Intelligence-is-Revolutionizing-Social-Media-to-Drive-Higher-Engagement-09-scaled.webp)

Adding to the conversation around deepfakes and artificial intelligence interventions, new research has explored the complexities of content moderation using AI and the related regulatory measures. . (Fletcher, 2018) suggested a multi-layered strategy that reduces false positives and increases the accuracy of deepfake detection by integrating AI algorithms with human monitoring. This hybrid methodology is in line with the advice of regulatory organisations like the European Union Agency for Cybersecurity (ENISA), which supports deepfake governance through a risk-based strategy (Gocen, 2023).

Furthermore, it has become clear that combining blockchain technology with AI might improve transparency and trust in digital material. In order to stop the spread of deepfake disinformation, Chan et al. (2020) investigated the viability of blockchain-based certification systems to confirm the legitimacy of media assets. This multidisciplinary approach highlights how blockchain technology and artificial intelligence may work together to fight issues related to digital manipulation.

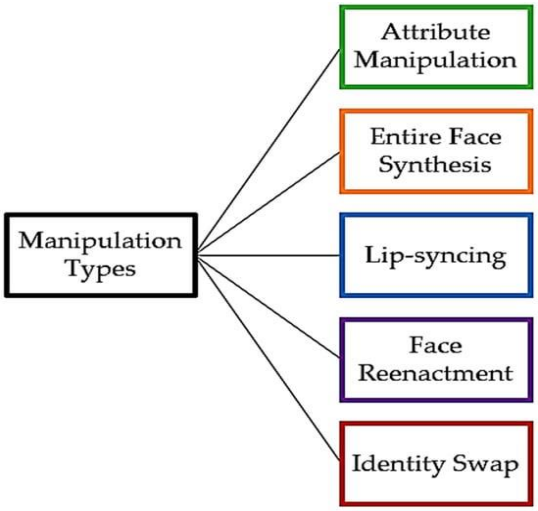

Fig 2.3: Deep Fake Manipulation Types*(* https://pub.mdpi-res.com/computers/computers-12-00216/article_deploy/html/images/computers-12-00216-g001-550.jpg?1698110664)

RESEARCH GAP

The literature review on deepfakes, AI interventions, and their influence on social networks identifies numerous areas where additional research might help advance understanding and close current gaps. Based on the literature evaluation, the following possible research gaps exist:

- Hybrid AI-Human Deepfake Detection: Look at models that combine AI and human verification for more accurate deepfake detection.
- Ethical Frameworks for Deep Fake Use: Investigate ethical concerns in deepfake development and transmission, including cultural and legal implications.
- Blockchain-AI Integration for Deepfake Verification: Investigate the feasibility and security of utilising blockchain to validate media authenticity in combatting deepfakes.
- User Perception and Behaviour: Examine how users perceive and interact with deepfakes to inform targeted interventions and education.
- Scalability and Fairness in AI Detection: Address scalability issues and biases in AI algorithms to provide fair and inclusive deepfake detection.
- Interdisciplinary Collaboration: Encourage collaboration among AI professionals, legal scholars, and policymakers to address deepfakes holistically.
- Long-Term socioeconomic Impacts: Investigate the socioeconomic and democratic implications of deepfake technology for media literacy and trust.

III.IMPACT OF DEEP FAKE ON SOCIAL NETWORKS

Deepfake technology's ascent has had a significant influence on social networks and ushered in a period of increased worry over false information, manipulation, and reliability. Artificial intelligence breakthroughs have led to a proliferation of deepfakes on digital platforms, which pose serious threats to the legitimacy and authenticity of online information. In addition to obfuscating the distinction between fact and fiction, this boom in manipulated media has accelerated the spread of false

narratives, swaying public opinion and undermining confidence in digital communication channels. This section explores the many impacts of deepfakes on social media platforms, backed up by pertinent data and academic analysis.

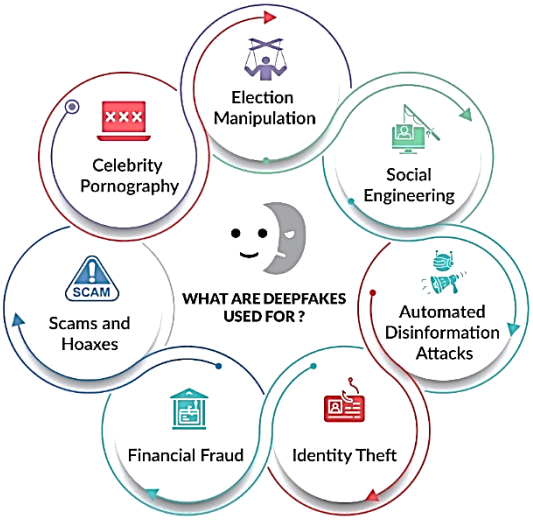

Fig 3.1: Deep Fake Impact *(* https://www.fortinet.com/content/fortinet-com/en_us/resources/cyberglossary/deepfake/_jcr_content/par/c05_container_copy_c/par/c28_image.img.jpg/1662490502931.jpg)

i. **Spreading False Information and Fake News** - On social media, deepfakes have played a part in the spread of false information and fake news. According to a Deeptrace analysis, there were more than 14,000 deepfake films uploaded online in 2019—a doubling from 2018 to 2019 (Shamanth et al, 2022) Deepfake material has become more prevalent, raising questions about the legitimacy and dependability of digital media platforms and making online information sources more scrutinised.

ii. **Controlling Public Opinion** - Deepfakes possess the ability to manipulate not just public opinion but also social media conversation. According to a Pew Research Centre poll, deepfake films significantly influence political discourse, according to 63% of participants (Sidoti et al, 2023). This impact highlights how social media users may be duped by false information and how deepfakes can change people's opinions.

iii. **Loss of Credibility and Trust** - The increasing occurrence of deepfakes has led to a progressive loss of credibility and confidence in digital media platforms. Sensity, a visual threat intelligence firm driven by artificial intelligence, said that between 2019 and 2021, the number of deepfake videos discovered increased by 330%, indicating a growing danger to internet trust (Passos et al, 2022). Social networks face difficulties preserving user confidence and thwarting misinformation operations as a result of this erosion of trust.

iv. **Effects on Privacy and Digital Identity** - Deepfakes are a serious threat to social network privacy and digital identity. Deepfakes may mimic people and fabricate stories by altering their appearances and voices. This can result in identity theft and damage to one's reputation. According to a NortonLifeLock survey, 76% of participants were worried about their personal privacy being compromised by deepfake technology (Muammar et al, 2023).

v. **Difficulties in Content Moderation** - The fast spread of deepfakes poses difficulties for social network platform administration and content regulation. Malicious deepfake material is difficult for platforms to identify and remove, which amplifies false narratives. The complexity of content moderation in light of developing deepfake technology was highlighted in a paper published by the Centre for Democracy & Technology (Helmus et al,2022).

IV. AI-BASED DEEPFAKE MITIGATION TECHNIQUES

As deepfake technology advances, the necessity for appropriate mitigation measures becomes critical in protecting social networks from the negative effects of synthetic media. Artificial intelligence (AI) is critical in creating novel methods to identify, refute, and remediate deepfake material. This section looks at several AI-based mitigation approaches that use complex algorithms and machine learning skills to counteract the spread of deepfakes on social networks.

1. Face Manipulation Detection Networks (FMDNs) for Detection

These are specialised AI models made to identify facial

modifications and manipulations, which are frequently linked to the creation of deepfakes. FMDNs are able to recognise modified or synthetic faces in multimedia information by analysing facial features, textures, and inconsistencies using deep learning architectures and algorithms.

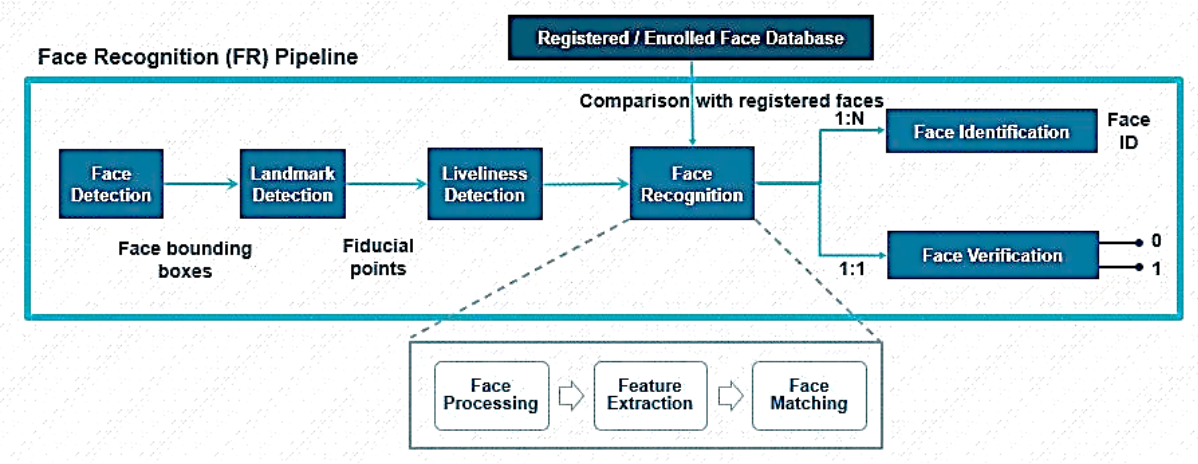

Fig 4.1: Facial manipulation recognition building block(https://www.eetasia.com/facial-recognition-fundamentals/)

PROS:

- **High Accuracy**: By utilising sophisticated deep learning algorithms and feature analysis, FMDNs show a high degree of accuracy when identifying modified faces.
- **Specificity:** These networks are useful for detecting deepfake content since they are designed to recognise minute changes and abnormalities unique to facial features.
- **Real-Time Detection**: FMDNs are appropriate for applications needing prompt reaction and intervention because they can detect altered faces in real-time.
- **Cross-Platform Integration**: FMDNs are compatible with a wide range of systems and platforms, such as social media sites, content moderation programmes, and forensic analysis tools.

CONS:

- **Data Dependency**: The quality and diversity of training data are critical to the efficacy of FMDNs, necessitating large-scale datasets covering a range of facial alterations.
- **Computing Resources**: Especially for large-scale or real-time detection applications, training and deploying FMDNs may require a substantial computing investment.
- **Domain-Specific tweaking**: To adjust to various deepfake generating methods and variations, FMDNs could need domain-specific tweaking and optimisation.
- **Adversarial Attacks**: Similar to other AI-based detection systems, FMDNs could be subject to adversarial attacks that take advantage of flaws in the network's architecture or training data to avoid detection.

APPLICATIONS:

- **Social Media Content Moderation**: In order to stop the spread of deepfake content, FMDNs are essential to the systems in place on social media platforms for detecting and flagging modified faces in photos and videos.
- **Educational Initiatives**: To inform users about the dangers of deepfake images and provide them with the ability to recognise modified face content, FMDNs can be integrated into media literacy courses and educational platforms.
- **Digital Identity Verification**: FMDNs are essential to digital identity verification systems because they guarantee the veracity of facial biometrics and stop identity fraud through the use of modified faces.
- **Media Integrity Verification**: To ensure the accuracy of facial imagery used in news stories, documentaries, and multimedia content, journalists, content creators, and media organisations use FMDNs.

2. Zero-Shot Learning and Few-Shot Learning

Zero-shot learning and few-shot learning are machine learning approaches that allow AI models to learn from little amounts of labelled data or even generalise to new classes in the absence of labelled instances. According to (Chen et al, 2023) in the context of deepfake identification, these strategies enable AI systems to recognise and categorise deepfakes with less training data, increasing flexibility and robustness to new and changing manipulation techniques.

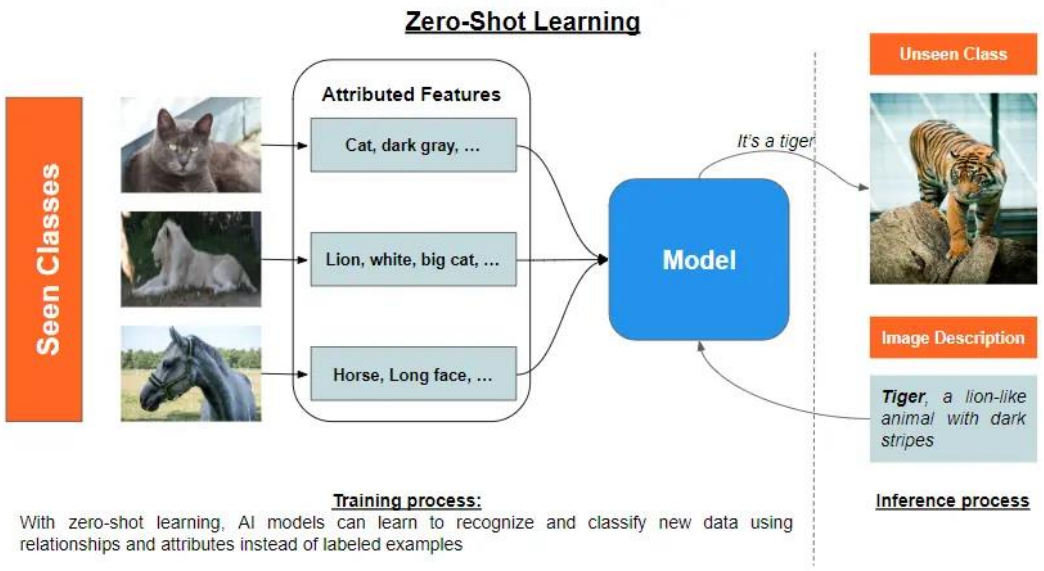

Fig 4.2: Zero-Shot Learning Scheme (https://saturncloud.io/images/blog/zero-shot-learning.webp)

PROS:

- **Generalisation**: Zero-shot and few-shot learning allow AI models to generalise to new deepfake variants or classes without explicit training, increasing flexibility and scalability.
- **Limited Data Requirement**: These approaches require little labelled data for training, making them appropriate for situations in which labelled datasets are limited or expensive to obtain.
- **Adaptability to Novel Classes**: AI models trained with zero-shot or few-shot learning may successfully categorise previously unknown deepfake classes, eliminating the requirement for ongoing retraining as new deepfake variations arise.
- **Reduced Bias**: By learning from a wide range of instances, zero-shot and few-shot learning algorithms can reduce biases in training datasets, resulting in more balanced and fair detection results.

CONS:

- **Model Complexity**: Using zero-shot and few-shot learning approaches may need complicated model structures and optimisation methodologies, which raises computational complexity and resource requirements.
- **Domain Specificity**: The efficiency of zero-shot and few-shot learning might differ amongst deepfake domains and scenarios, necessitating domain-specific tweaking and adaptation.
- **Performance Dependence on Training Data Quality:** The quality and diversity of initial training data have a substantial influence on the performance and generalizability of zero-shot and few-shot learning models.
- **Interpretability Challenges**: The underlying workings of zero-shot and few-shot learning models can be complicated and difficult to understand,
- making it difficult to explain detection choices and outcomes.

APPLICATIONS :

- **Quick Deployment Systems**: Zero-shot and few-shot learning models are appropriate for quick deployment in cases requiring immediate identification of new deepfake variations, such as breaking news or viral content outbreaks.
- **Resource-Constrained Environments**: When large labelled datasets are unavailable, zero-shot and few-shot learning algorithms provide a feasible alternative for developing strong deepfake detection systems.
- **Continuous Learning Platforms**: Platforms and systems that need to learn and adapt to developing deepfake threats might benefit from zero-shot and few-shot learning's incremental learning capabilities.
- **Cross-Domain Detection**: Zero-shot and few-shot learning models may generalise across many deepfake domains, such as pictures, videos, and audio, making them appropriate for multi-modal deepfake detection applications.

3. Multimodal Analysis

Multimodal analysis is a machine learning method that uses many modalities, such as text, picture, audio, and video, to improve detection skills(Lomnitz et al, 2020). In the context of deepfake detection, multimodal analysis enables AI systems to catch subtle details and

inconsistencies across many media formats, hence enhancing detection algorithms' accuracy and reliability.

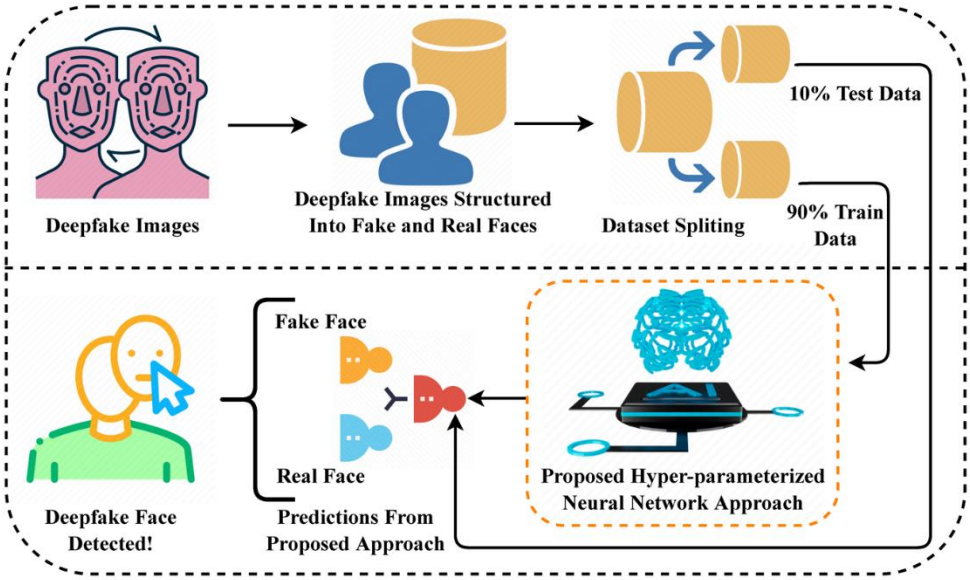

Fig 4.3: Multimodal Analysis Scheme (https://www.mdpi.com/applsci/applsci-12-09820/article_deploy/html/images/applsci-12-09820-g001.png)

PROS:

- **Complete Analysis**: Multimodal analysis allows for a complete study of deepfake content by taking into account numerous modalities at the same time, uncovering tiny abnormalities and inconsistencies that would otherwise go unnoticed.
- **Improved Accuracy**: By combining data from several sources, multimodal analysis improves detection accuracy and resilience while lowering false positives and false negatives in deepfake detection.
- **Contextual Understanding**: Analysing several modalities gives context and semantic understanding, allowing AI systems to distinguish between authentic and modified material based on contextual cues and relationships.
- **Adaptability to Evolving approaches**: Multimodal analysis approaches may adjust to new deepfake generating techniques and variants, making them resistant to changing manipulation strategies.

CONS:

- **Complexity**: Implementing multimodal analysis may be difficult, necessitating advanced algorithms, feature extraction approaches, and model architectures to successfully integrate data from many modalities.
- **Data Fusion issues**: Combining information from several modalities necessitates careful data fusion approaches to avoid information loss or noise, which presents issues in data pre-treatment and feature selection.
- **Computational Resources**: Multimodal analysis might need substantial computational resources and processing capacity, especially for real-time or large-scale deepfake detection applications.
- **Interpretability**: The interpretability of multimodal analytic results can be difficult, since the integration of many modalities might obfuscate the underlying decision-making process, restricting explanation.

APPLICATIONS:

- **Video Hosting Platforms**: Major video hosting platforms may utilise multimodal analysis to detect deepfake movies by examining audio-visual material, text descriptions, and user interaction patterns.
- **News Verification and Fact-Checking**: Multimodal analysis may be used by media organisations and fact-checking agencies to ensure the validity of news items by analysing multimedia material, metadata, and contextual information.
- **Social Media Content Moderation**: Social media platforms may include multimodal analysis techniques into their content moderation systems to detect and flag deepfake content in a variety of formats, including photos, videos, and text.
- **Cybersecurity and Fraud Detection**: Multimodal analysis may be used to detect deepfake-based cyberattacks, phishing attempts, and identity theft schemes by combining text, audio, and visual indicators.

4. Domain-Specific Networks for fake detection

They are specialised artificial intelligence models that are made to identify fraudulent information in certain domains, including pictures, videos, or audio files. These networks look for anomalies suggestive of fabricated or altered content using domain-specific features, patterns, and traits. Domain-Specific Networks improve detection accuracy and reliability by being specifically designed to perform well in their assigned domain. their categorization. (Zhang et al, 2016).

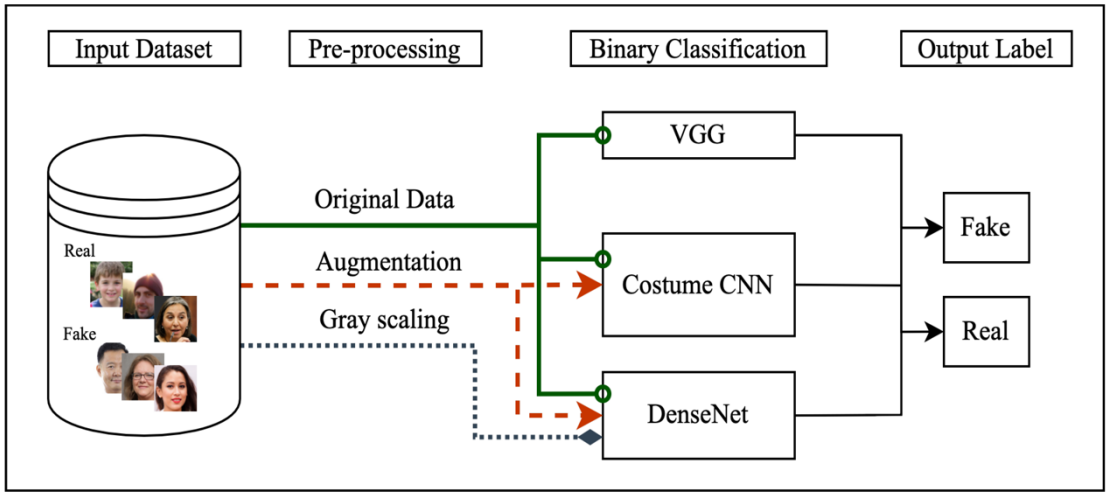

Fig 4.4: Domain Specific Networks (https://www.mdpi.com/jcp/jcp-02-00007/article_deploy/html/images/jcp-02-00007-g001.png)

PROS:

- **Great Accuracy**: By utilising domain-specific features and patterns for accurate identification, Domain-Specific Networks show great accuracy in identifying fraudulent information inside their assigned domain.
- **Optimised Performance**: These networks are tuned for their particular domain, which produces effective and efficient detection methods that are specific to the domain's peculiarities.
- **Reduced False Positives**: Domain-Specific Networks can minimise the misclassification of real information as fraudulent by concentrating on domain-specific attributes.
- **Real-Time Detection**: A few Domain-Specific Networks have this feature, which makes them appropriate for applications that need to be addressed and responded to right away.

CONS:

- **Limited Scope:** Domain-Specific Networks must use various models or algorithms for different domains, such as photos, videos, or audio, in order to detect fraudulent information outside of their assigned domain.
- **Data Dependency:** area-Specific Networks are only as effective when their training and validation datasets are large and varied within the area in question.
- **Generalisation Challenges:** These networks may find it difficult to generalise across various fake content variants or methodologies within the domain, necessitating ongoing adaptation and improvement.
- **Domain-Specific Tuning**: Skill and iterative improvement may be needed to fine-tune and optimise Domain-Specific Networks for optimal performance within the assigned domain.

APPLICATIONS:

- **Image Forgery Detection:** Domain-Specific Networks are highly effective in identifying picture alterations, including content change, image splicing, and retouching.
- **Video Deepfake Detection**: These networks play a crucial role in the realm of video content by detecting deepfake movies and differentiating between real and altered visual material.
- **Audio Manipulation Detection:** To protect the integrity of audio recordings, domain-specific networks are capable of identifying audio manipulations such voice cloning, audio deepfakes, and synthetic speech.
- **Evidence Analysis:** Using Domain-Specific Networks, law enforcement organisations can do forensic analysis to identify fraudulent content and verify the legitimacy of digital evidence found within particular domains.

V.DISSCUSSION

The spread of deepfake technology has had a major

influence on social networks, bringing substantial difficulties in terms of disinformation, privacy intrusions, and trust loss. In this context, the discussion centres on the consequences of deepfake impact and the efficacy of mitigating approaches, such as AI-based solutions, in resolving these issues.

The impact of deepfake technology:

Deepfakes have developed as an effective tool for altering digital information, including movies, photos, and audio recordings, with remarkable realism.(Helmus, 2022)The consequences of deepfake technology extend across several domains:

- **Misinformation and Disinformation**: Deepfakes contribute to the spread of misinformation and disinformation, resulting in public confusion, polarised attitudes, and diminished confidence in media sources.
- **Social and Political Manipulation**: Deepfakes may be used to undermine democratic processes, distribute misinformation, and shape public opinion.
- **Privacy problems:** The production and transmission of deepfakes create major privacy problems since people's faces, voices, and identities can be impersonated without their permission, resulting in reputational harm and privacy violations.
- **Technical Arms Race:** The fast growth of deepfake methods needs ongoing adaptation and innovation in detection and mitigation tactics, resulting in a technical arms race between creators and defenders.

The Effectiveness of Mitigation Techniques:

AI-based mitigation approaches are critical in reducing the impact of deepfakes. These methods use machine learning algorithms, data analysis, and pattern recognition to detect and minimise synthetic media manipulation. The main points of debate include:

- Detection Accuracy: AI-powered detection systems have demonstrated promising results in reliably recognising deepfake material, lowering false positives, and increasing overall detection rates.
- Scalability and Real-Time Detection: AI models' scalability enables real-time deepfake detection, allowing platforms to discover and mitigate altered material before it becomes viral.
- Interpretability and Transparency: Domain-specific network approaches offer clear explanations for detection conclusions, which improves confidence and accountability in deepfake detection systems.
- Cross-Modal Analysis: Combining various data modalities improves detection accuracy and resilience by catching subtle traits and discrepancies across media types.
- Privacy-Preserving Techniques: Federated learning and privacy-preserving AI technologies protect users' privacy while enhancing detection skills through collaborative learning and data sharing.

VI.CONCLUSION AND FUTURE SCOPE

Deepfake technology has become a double-edged sword in the digital era, threatening online authenticity but also providing unique chances for artistic expression. Deepfakes have a wider influence on social networks, including disinformation transmission, political manipulation, and privacy breaches, necessitating strong mitigation techniques.

AI-based mitigation strategies have emerged as a key tool in addressing the negative consequences of deepfakes. These approaches, which range from deepfake detection algorithms to Face Manipulation Detection Networks (FMDNs) and multimodal analysis, exhibit excellent accuracy, scalability, and real-time capabilities. Regardless of its success, further study is required to address issues such as model interpretability and data sensitivity.

Real-world deployments of AI-powered content moderation systems have had a noticeable impact on reducing the propagation of deepfake content on social media platforms. Privacy-preserving solutions included into these systems solve user privacy issues, resulting in increased platform integrity and user confidence.

Future research should focus on improving model interpretability, encouraging multidisciplinary partnerships, and boosting media literacy to help viewers discern between real and manipulated material. Ethical issues and ongoing innovation in AI technologies are critical to successfully navigate the changing world of deepfake dangers.

REFERENCES

[1] Kietzmann, J., Lee, L.W., McCarthy, I.P. and Kietzmann, T.C., 2020. Deepfakes: Trick or treat?. *Business Horizons*, *63*(2), pp.135-146.

[2] Montasari, R., 2024. The Dual Role of Artificial Intelligence in Online Disinformation: A Critical Analysis. In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (pp. 229-240). Cham: Springer International Publishing.

[3] Stroebel, L., Llewellyn, M., Hartley, T., Ip, T.S. and Ahmed, M., 2023. A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology*, *7*(2), pp.83-113.

[4] Montasari, R., 2024. Responding to Deepfake Challenges in the United Kingdom: Legal and Technical Insights with Recommendations. In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (pp. 241-258). Cham: Springer International Publishing.

[5] Ferrara, E., 2024. GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, pp.1-21.

[6] Helmus, T.C., 2022. Artificial Intelligence, Deepfakes, and Disinformation.

[7] Wagner, T.L. and Blewer, A., 2019. "The word real is no longer real": Deepfakes, gender, and the challenges of ai-altered video. *Open Information Science*, *3*(1), pp.32-46.

[8] Fletcher, J., 2018. Deepfakes, Artificial Intelligence, and Some Kind of Dystopia. *Theatre Journal*, *70*(4), pp.455-471.

[9] Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dute, I., Khan, S. and Parkinson, S., 2023. A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access*.

[10] Köbis, N.C., Doležalová, B. and Soraperra, I., 2021. Fooled twice: People cannot detect deepfakes but think they can. *Iscience*, *24*(11).

[11] Kim, B., Xiong, A., Lee, D. and Han, K., 2021. A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PloS one*, *16*(12), p.e0260080.

[12] Li, M. and Wan, Y., 2023. Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Research*, *33*(5), pp.1750-1773.

[13] Ahmed, S.R., Sonuç, E., Ahmed, M.R. and Duru, A.D., 2022, June. Analysis survey on deepfake detection and recognition with convolutional neural networks. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 1-7). IEEE.

[14] Remya Revi, K., Vidya, K.R. and Wilscy, M., 2021. Detection of deepfake images created using generative adversarial networks: A review. In *Second International Conference on Networks and Advances in Computational Technologies: NetACT 19* (pp. 25-35). Springer International Publishing.

[15] Rahman, A., Islam, M.M., Moon, M.J., Tasnim, T., Siddique, N., Shahiduzzaman, M. and Ahmed, S., 2022. A qualitative survey on deep learning based deep fake video creation and detection method. *Aust. J. Eng. Innov. Technol*, *4*(1), pp.13-26.

[16] Gocen, I., 2023. *European Union's Approach to Artificial Intelligence in the Context of Human Rights* (Doctoral dissertation, Dokuz Eylul Universitesi (Turkey)).

[17] Barman, D., Guo, Z. and Conlan, O., 2024. The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. *Machine Learning with Applications*, p.100545.

[18] Chan, C.C.K., Kumar, V., Delaney, S. and Gochoo, M., 2020, September. Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media. In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)* (pp. 55-62). IEEE.

[19] Shamanth, M., Mathias, R. and MN, D.V., 2022. Detection of fake faces in videos. *arXiv preprint arXiv:2201.12051*.

[20] Sidoti, O. and Vogels, E.A., 2023. What Americans Know About Al, Cybersecurity and Big Tech.

[21] Passos, L.A., Jodas, D., Costa, K.A., Souza Júnior, L.A., Rodrigues, D., Del Ser, J., Camacho, D. and Papa, J.P., 2022. A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, p.e13570.

[22] Muammar, S., Shehada, D. and Mansoor, W., 2023. Digital Risk Assessment Framework for Individuals: Analysis and Recommendations. *IEEE Access*.

[23] Chen, J., Geng, Y., Chen, Z., Pan, J.Z., He, Y., Zhang, W., Horrocks, I. and Chen, H., 2023. Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey. *Proceedings of the IEEE*.

[24] Lomnitz, M., Hampel-Arias, Z., Sandesara, V. and Hu, S., 2020, October. Multimodal approach for deepfake detection. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)* (pp. 1-9). IEEE.

[25] Zhang, J., Li, J., Li, X.L., Shi, Y., Li, J. and Wang, Z., 2016. Domain-specific entity linking via fake named entity detection. In *Database Systems for Advanced Applications: 21st International Conference, DASFAA 2016, Dallas, TX, USA, April 16-19, 2016, Proceedings, Part I 21* (pp. 101-116). Springer International Publishing.

[26] Sánchez-Junquera, J., Villaseñor-Pineda, L., Montes-y-Gómez, M., Rosso, P. and Stamatatos, E., 2020. Masking domain-specific information for cross-domain deception detection. *Pattern Recognition Letters*, *135*, pp.122-130.