

Analysis of Cybersecurity : Deepfake and misinformation detection :role of AI and ML in curbing the above

Bhavika Khanna

Vasant Valley School

DOI: 10.46609/IJSSER.2026.v11i10.025 URL: <https://doi.org/10.46609/IJSSER.2026.v11i10.025>

Received: 2 October 2026 / Accepted: 15 October 2026 / Published: 20 October 2026

ABSTRACT

The paper attempts to address the increasing spread of Deepfake's being used with the help of AI and ML technology to spread crime in the financial set up, government, politics, entertainment industry as well as in the lives of ordinary citizens by keeping them under 'digital arrest'. The same technology that was developed for the convenience and for the benefit of it's citizens is being used in a negative manner. Countries have developed expensive safeguards that have been embedded in the systems to thwart any malafide attack. The expenditure in this area is necessary to prevent harm to the people of a country. The paper address the timeline of the increasing penetration of Deepfake as well as the technology being developed to prevent it.

Key Words: Social media platforms, Deepfake, Cybercrime, Cyber security, misinformation, false information, Deep learning, Cyber Warfare.

Introduction

The rising awareness of social media platforms and its widespread usage as benefited society but has also brought in it's wake a plethora of negative usage that ranges from the spread of both misinformation as well as false information. This has spread across all sectors and sections of society that range from social, political economic, defense and the dark underbelly of crime. This technology is used to spread false information, encouraging hatred, exploiting the public and the government as well as financial exploitation.

The world has reached such a stage that given the extent of incorrect information that is being spread it becomes necessary to distinguish between 'false information' and 'misinformation'.

Misinformation is primarily false information that is being shared, but not directly to deceive people, but disinformation is a process by which false information is being spread purposely to

manipulate and mislead people. Though the intent may be different, both cause a large amount of harm to the citizens of the world.

Deepfake is an amalgamation of the words 'deeplearning', and 'fake'. These are video's, images and audio, that have been generated by Artificial Intelligence, AI tools, and audio-video editing software. They have the ability to copy both 'real', and 'fictional', people and creating them in any type of surroundings that they wish. In the earlier stages, deepfakes were recognised by their unnatural movements, artificial facial reactions, changes during the presentation of audio quality, mismatches in colour and lighting. But the one's in use today is sophisticated. Difficult to recognise with the naked eye. To differentiate between the 'real' and 'fake' require the use of sophisticated technology.

Deepfakes has been used constructively to develop opportunities in education and entertainment. This technology is one of the most controversial ones as it has led to a 'revolution', as well as a 'threat', to society. It is a true 'double edged' sword as it raises issues with regard to it being misused in misinformation campaigns, identity fraud and cyberwarfare.

Initially it was mainly used as a tool in the entertainment industry where it was used to make realistic characters in video games, or even to resurrect deceased characters in Films. Often if the actor passed away during the making of a film, then instead of changing or rewriting the character, AI technology was used to resurrect the actor, and complete the film.

The increasing use of this technology by people spreading mis information and deliberately misleading and conveniently falsifying information as led to a trust deficit in social media posts. This has happened due to the exponential growth of social media, that has primarily occurred due to reaping the gains of economies of scale, thus reducing the cost of usage of expensive technology. This widespread usage has been made possible due the availability of inexpensive smart phones, wide spread internet facilities and the prevalence of sophisticated technology. *This paper studies the role of this negative use of technology and the corresponding cyber security laws that need and are being implemented to curb it's spread. What is being attempted by countries to keep pace with the negative result that this technology has presented in a continuous?*

Background

The growing spread of misinformation and fake news has been possible since the emergence of digital visual media. Earlier such type of news required specialised expertise and time consuming tools like Adobe 'Photoshop' or 'lightroom', but with the growth of technology and advancement in AI, videos and images are easily altered or tampered with, resulting in extremely realistic forgeries that are difficult to distinguish from the original.

The accessibility to AI tools and software have become easily available to all. The availability of this technology has made it easy for those with limited knowledge to gain access and use it in the manner they want to. This technology has been used to create non consensual information that are able to hurt eminent personalities and either enhance or spoil their reputation. These are actively used by political parties to strengthen their cause and/or damage that of the opposition. This technology increasingly relies on what is called “Generative Adversarial Network”, (GAN’s”) (Priyanshi Oz, Nirjharaa Patel, Ayushi Patel.2020). It is used by firms to spread “fake communication”, maybe about themselves or their rivals to obtain advantages, that is likely to be reflected in their share price, sales and revenue. There is an imperative need for it to be addressed as it creates a negative and far reaching impact on various citizens in a country. It has both societal and ethical consequences, impacting public trust and social cohesion.

Research has indicated that Deepfakes are AI “generated digital content, normally a video, audio, or an image, which is manipulated by using Deep Learning(DL) algorithms. This technology alters the original matter, replaces it with a false one in such a manner that the so called ‘new’, one looks to be the ‘correct and the true one’.(Zhang et al) The tools that are adopted are in the form of ‘ visual disinformation technique by creating a hyperrealistic synthetic media”. It is the use of Generative Artificial Intelligence (GenAI), and Machine Learning(ML) and Deep learning (DL), that use Autoencoders(AE) and Generative Adversarial Networks (GAN’s) (Kaur .,et al)

The use of this technology has increased by 700% in 2023 in the Fin Tech sector. In 2024 a CEO in the US was manipulated to transfer \$25.6 million due to a request by an AI generated video call. This is one example of the many such examples that are being reported on a regular basis (Alakananda Mitra-2024). Besides Finance this technology is being used even in political arena. An example being the fabricated videos of Ukrainian President Zelenskyy announcing surrender and manipulated footage of former U.S. President Barack Obama delivering controversial messages (Saraju P. Mohanty.,2024)

Given the wide range in which Deep Fake is being adopted across “privacy violations, identity theft, scams, misinformation, political manipulation, hate speech, defamation, and large-scale societal disruption”, it has ultimately threatened the ethos of trust, information integrity, and global security” (Nachaat Mahmood- 2024).

Given the serious implications of this technology, it is important to draw a timeline of how it became a formidable force that it is today. Most deepfake content is visual and focuses on human faces, this was the first milestone that this technology reached. Subsequently other techniques were developed like voice swapping, text-to-speech, natural sounding speech and AI text generation, lip syncing (Nachaat Mahmood- 2024). All of them are considered extremely

important in spreading false and fake narratives. Given that world is primarily always watching and absorbing on line content, for them the fake news that is being broadcast seems real, regardless of it's form and where it is coming from.

Reviewers and researchers have worked with large data sets to understand and work towards dis aggregating, the generation of information. For this various techniques are reviewed across audio, video, and images. Fake news both textual and multimedia based is absorbing this content at a fast pace. In this process both synthetic and real data is analysed, by taking into account the properties, use of the cases and their limitations. Evaluation of data across a large number of studies, have resulted in the evaluation of various performance metric, which are extracted from the data, which is then compared across a number of data sets to understand and recognise the fact that the content is real or fake. It becomes essential to categorise the fake content, how is it possible to detect it, what can be done to prevent it given the resources at hand.

Research Gap and Methodology

This paper would analyse the importance of methods and technology that would control the spread of fake news, and Deepfake technology. What is the role that cyber security plays in controlling this menace? How can the citizens of a country be made aware to distinguish between fake and the real content? How important is it to use the correct AI and ML technology to control this menace?

The methodology adopted would primarily be of the mixed method type that would use qualitative and quantitative data from authentic secondary sources.

Analysis

Table 1; Analysis of Percentage of Fake News in India in 2024 in different Areas.

Fake news	percentage
Political	46%*
Financial (AI-generated misinformation overall)	8.35%**
Health	11.1%**
Religious	16.8%**
International Affairs	15%**
Education	NEET, UPSC
General Information	33.3%**

Source: * ISB/CyberPeace, **Boom Fact Check

The type of misinformation in different areas is listed below.

Politics: False or misleading statements, speeches, photos and videos attributed to politicians, False claims about election results, voting, polling and election events, Claims that EVMs were hacked, votes were manipulated or machines were being misused

Health: This constitutes false claims about medicines, vaccines, diseases, treatments and health products.

Religious: False claims involving various religions, religious practices, religious leaders and events.

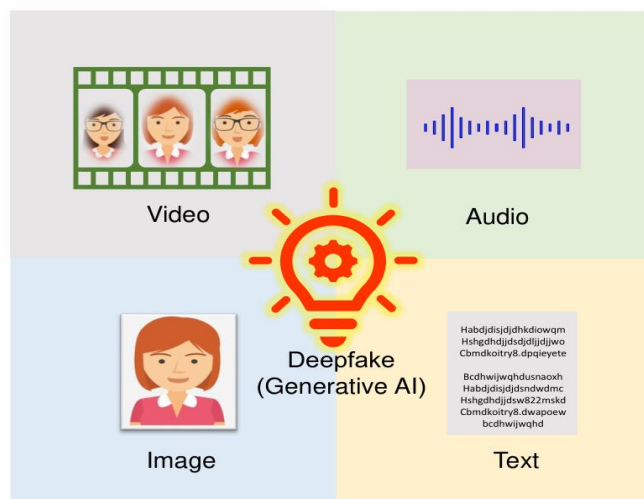
International Affairs: False claims about wars, foreign governments, international organisations and events abroad

Education: Fake information about examinations, results, universities, entrance tests, admissions and educational policies.

General Information: Fake information about examinations, results, universities, entrance tests, admissions and educational policies, Misleading claims about communities, society or public behaviour

The present status of advance that exists in Deepfake technology is due to the use of approaches like ‘Federated learning’, ‘Blockchain’, and ‘Explained AI’. This involves algorithmic development, datasets, tools, in a unified setup.

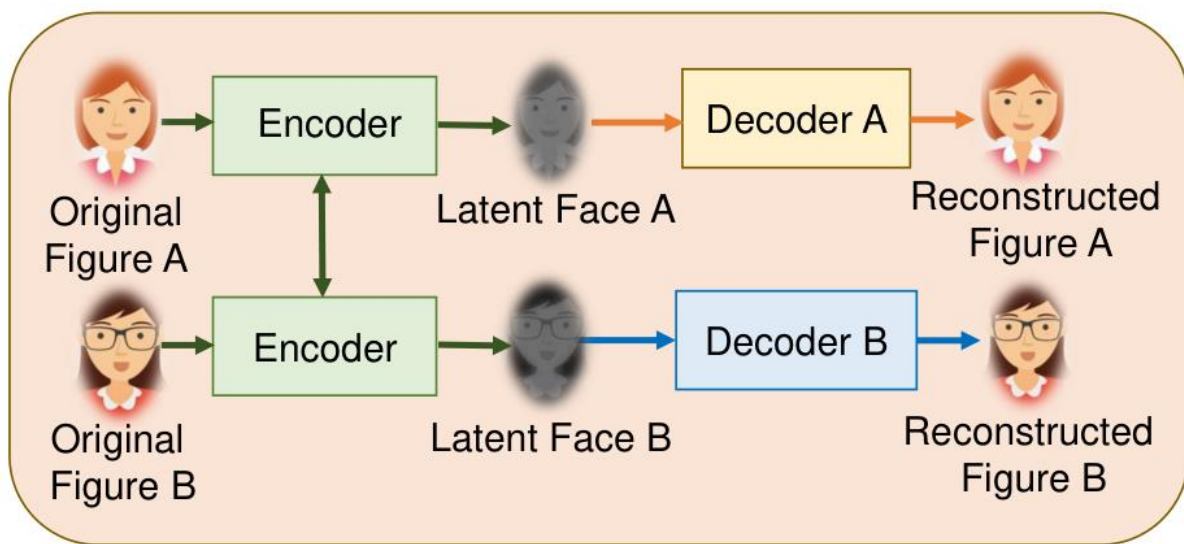
Image 1: Creation of Deep fake videos with the required audio



Source: Nachaat Mahmood- 2024

These are generated by using both deep learning and supervised learning models. The two main types of models use. These autoencoders are 'variational autoencoders' and 'Generative Adversarial Networks (GANs)'. For facial recognition 'variational autoencoder' is used, where they compress and encode input data to a lower dimensional latent space, which is then reconstructed to provide output data in this case the image.'Face swap' is possible by using the 'autoencoders'.

Image 2: Using 'autoencoder', to create and swap images.



Source: Nachaat Mahmood- 2024

'GANs' are used to create realistic images. The 'generator', continues to reconstruct images till the 'discriminator', assesses the quality. If the generator fails the 'discriminator's', test, then it is ordered to continue to create till it can actually 'fool the discriminator'. The tools that are used to create these authentic deepfakes are: FaceApp, DeepBrain, ,DeepSwap, Reface, FaceSwap, MorphMe, DeepFaceLab. Chat GPT then combines with DeepBrain to produce real time conversation.

The discovery of 'ChatGPT', has led to add authentic voice to sophisticated images, videos, and texts. The superiority of the technology is so immense that it is possible for the person created as Deepfake to carry out an intelligent conversation. It is the sophisticated GPT-4LLM that is used as it accepts both text as well as image as inputs. The depth of the technology is so immense that all one needs is a prompt of the content to be given to ChatGPT, and it manages to create a credible outline dialogue. This can be done in any language and voice. The time taken to produce these videos has also been reduced. Thus with this technology, it has reduced barriers of entry

into the ‘world of crime’, becoming cost effective without any outside professional help. There are start ups like Hour One, Synthesia and Uneeq that have integrated ChatGPT, Large Language Models (LLMs), generative AI and other AI that is equipped with the ability to make and create “anything from nothing”.

Growing Demand For Cyber Security

With the growing infiltration of Deepfakes, it becomes extremely important and necessary to construct, develop, and adopt alternative technology to adequately address this threat. The exponential growth of the digital revolution has led to a revolution in the way everyone communicates, does business, and receives information. The speed at which various tasks are performed in any section of society indicates the manner in which ‘internet’, has revolutionised the whole world. Advances in information and technology, along with the speed at which it is obtainable has led to the fact that it is ‘indispensable. But this has led to numerous security challenges, to all stake holder, namely individuals, businessmen, financiers, and the government.

Cyber security, is the protection of organisations as well as their Information Technology Systems from Cyberattacks. This security tries and preserves confidentiality, integrity, as well as the stored information that exists in the systems

Among the various important factors that are involved, the first is the ‘Basic Power System’, which has to be provided on a continuous basis for all end users. Power Electronics play an extremely important role. There have to be systems in place that ensure ‘back up’, in case there are fluctuations or complete shut down of the ‘power grid’.

Due to the exponential growth in the Digital World, there are a number of security challenges that have emerged. This has led to the continuous employment of cybersecurity measures, legal framework as well as education and awareness campaigns to warn citizens of the various types of mala fide cyber attacks that are in existence and that are being developed. The rise of cyber attacks have increased manifold with the increasing use of cloud computing, Internet of Things (IOT), and 5G technology. The widespread use of these facilities prevent the traditional security measures from performing their jobs efficiently.

Cyber security instruments are used in every sphere from small enterprises, government agencies, military groups, health care providers, educational institutions, energy suppliers, and transportation systems. Their primary aim being to protect personal data, secure vital infrastructures and guarantee the confidentiality and integrity of sensitive information.

To address this extremely imminent threat it is important to integrate artificial intelligence (AI), and machine learning (ML) within the security systems that are existing to prevent cyber crimes.

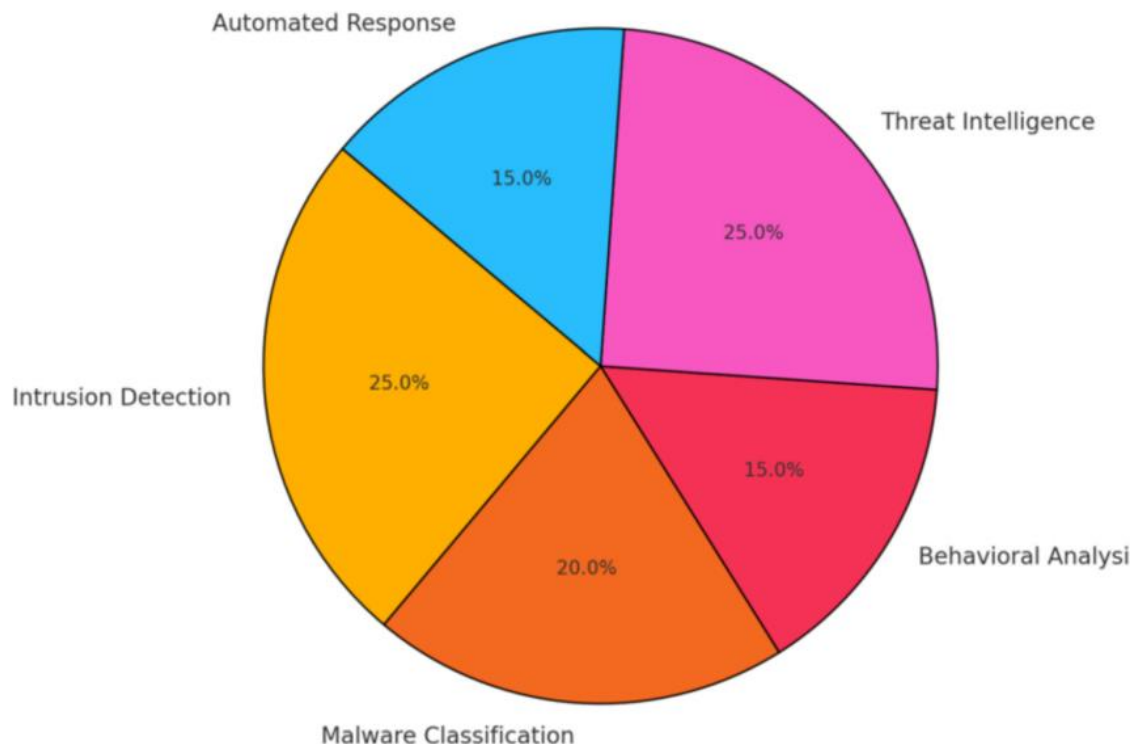
These systems have to be geared to detect, respond, and completely nullify complex cyber attacks. The use of AI and ML can further automate systems to detect “intrusion detection”, “malware analysis” and “vulnerability assessment.”

The latest type of technology that is being used for ‘cyber crimes’, as well as ‘cyber prevention’, is the rise of AI(XAI), which is the amalgamation of AI with quantum computing, along with the collaboration of cyber defense, keeping privacy intact. But inspite of these developments traditional AI driven security solutions still have to deal with ‘zero day attack detection’ and ‘adversarial machine learning (AML)’, where attackers manipulate AI models by injecting inputs that lead to mis classification and manage to bypass security barriers.

Approaches to address Cyber Attack

The diagram below is an example of how Cyber attacks can be prevented :

Diagram 1: AI and ML applications by Cyber Security



Source: Google images

The problem areas are ‘adversarial AI, automated threat intelligence, and to address it through AI driven security orchestration. The above diagram gives some idea on how attacks on cyber security can be addressed.

Other ways that these attacks are warded of are by deploying solutions to multistage attack (MSA), APT, DOS, wireless injection attacks, botnets and malicious activities.

These attacks can be prevented by:

- The Machine Learning Approach; this has helped in identification of an intruder in the system as well as it's precise location.
- Supervised Learning Approach; This occurs by injecting a machine learning algorithm into the particular instances through “C4.5 and ID3 algorithms, Artificial Neural Network, Hidden Markov Model (HMM), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Naïve Bayes”.(Sangeetha Prabhu, et al, (2020);
- UNSUPERVISED LEARNING APPROACH; this approach helps in distinguishing between normal results and those that have been tampered with.
- Semi supervised learning approach; this increased the expenses involved for a hacker to attack the system.
- Reinforcement Learning Approach; this helps in truncating the attack, but to be completely effective it has to be combined with other approaches.
- Detection by Genetic Algorithm Approach; This leverages network traffic through statistical capabilities to detect any unusual activity present in the traffic stream, and can detect novel attacks and activities.

Individual countries are using their own methods in trying to control the rampant spread of Cyber Crime. In the US, a bill (HF 1370) was passed in the Minnesota House that sharing of deepfake pornography is a criminal offense, but it was not passed by a majority vote. In India Cyber Crimes have been announced as a new “threat to Democracy” (Mr. Narendra Modi., 2024). A new watchdog website has been announced; NCRP (National Cyber Crime Reporting Portal, CRAU (Cyber Rapid Action Unit) and WebD (an artificial intelligence and process automation watchdog platform to track illegal URLs, gather digital evidence and automatically submit takedown orders).

Companies to have upped the ante, Sony has developed an in camera authenticity technology to address Deepfakes, by introducing digital signature thus authenticating the real image. Other

areas that technology is working is the development of digital watermarking and distributed ledgers.

There has been a shift to a more proactive approach rather than a reactionary one. AI and ML have transformed cybersecurity into resilient intelligent systems. These models by learning from new processes and data have managed to evolve to prevent and recognise the most sophisticated attacks, and have managed to set up a defense of a number of layered security devices. These techniques play an important role in threat hunting, malware analysis, insider threat detection among the many attacks, all to stay ahead of cyber attackers. The advanced technology adopted by countries today have become a necessity, they are no longer considered a luxury.

Limitations of the Study

The main limitations of the study is that the paper was dependent on authentic secondary data. It was not possible to conduct interviews or surveys to authenticate through primary research what has been observed through the secondary sources. This was mainly due to lack of time and finances.

Conclusion

Cyber crime and Deepfake are the most disturbing and evil trends that have developed with the increasing development of AI and ML. Ideally this sophisticated technology should have been just used for the development of mankind, and their development, but currently the same technology is being used to harm citizens. This not to say that the beneficial reach of this technology is not being witnessed, but there is a 'dark' side to it too, which needs to be urgently addressed and curbed. Hackers have managed with this technology and with increased digitilisation to produce fake videos, hold senior and dependent citizens to digital arrest, demanding huge sums of money from the vulnerable by faking their identity, and pretending to be government officials. This is also being used in political fields as well as in the entertainment industry. AI and ML development have to be one step ahead to be able to curb this crime. Governments individually have set up laws and procedures to warn their citizens as well as imposing hefty fines and imprisonment. The best way to address this issue is to create awareness and educate the people against falling in their trap. As far as companies are concerned, a large number of layers of security have been put in place to prevent hackers from entering into the system.

Bibliography

Admass, W. S., Munaye, Y. Y., & Diro, A. A. (2024). Cyber security: State of the art, challenges and future directions. *Cyber Security and Applications*, 2, 100031. <https://doi.org/10.1016/j.csa.2023.100031>

Caseiro, L. M. A. (2023). Power system resilience (end user) and critical infrastructure. Science Direct. Elsevier.

Jaggi, K. (2025). *Advancing cybersecurity strategies: Balancing threat detection, compliance, and resilient architectures*. SSRN. <https://doi.org/10.2139/ssrn.5124287>

Liaropoulos, A. N. (2015). A human-centric approach to cybersecurity: Securing the human in the era of cyberphobia. *Journal of Information Warfare*, 14(4), 15–24.

Mahmood, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Springer, 6969–7055.

Mandal, U., Setua, S. K., & Sarma, S. S. (2026). Deepfakes and beyond in the era of AI-generated disinformation: A systematic literature review of emerging technologies, challenges, and policy recommendations. Science Direct.

Mitra, A., Mohanty, S. P., & Kougianos, E. (2024). *The world of generative AI: Deepfakes and large language models*. Science Direct.

Schiliro, F. (2023). *Towards a contemporary definition of cybersecurity*. arXiv. <https://doi.org/10.48550/arXiv.2302.02274>., Research Gate