AI-Powered Detection of Deepfakes and Misinformation on Social Media

Kshithij Sangam[1], Bhavisha[2], Shameeksha[3], Spoorthi[4]
[1,2,3,4]*Dept. of Master of Computer Applications, Shree Devi Institute of Technology, Kenjar, Mangaluru*

***Abstract*—The rapid rise of AI has enabled the creation of realistic deepfakes and the widespread circulation of false information on social media. While these trends raise concerns around privacy, security, and public trust, AI also provides a path toward automated detection and mitigation. The study looks at different ways to spot manipulated images, altered videos, and misleading text by applying techniques from deep learning, NLP, and computer vision. It also reviews the datasets and detection models used, evaluates how well they perform, and highlights both the practical difficulties and ethical questions that should guide future work.**

***Index Terms*—Artificial intelligence, Deepfake Detection, Misin- formation, Natural Language Processing, Social Media Analysis, Computer Vision**

I. INTRODUCTION

The rapid expansion of digital communication has transformed social media into one of the most influential platforms for sharing information globally. While these networks empower individuals to exchange knowledge and opinions, they also serve as fertile ground for the spread of mis- leading and manipulated content. Among the most pressing concerns are *deepfakes*—synthetic media generated using artificial intelligence to convincingly mimic real individuals—and *misinformation*, which involves the deliberate or accidental distribution of false narratives. Both threaten the integrity of public discourse, erode trust in institutions, and in extreme cases, destabilize democratic processes and social harmony.

Deepfake technology has grown swiftly through progress and machine learning, beginning with generative adversarial networks (GANs) and later diffusion models. While these tools were first adopted in film, entertainment, and digital art, they are now frequently misused for damaging purposes. Exam- ples include fabricating political speeches, producing non- consensual explicit material, and manipulating digital evidence in legal settings.

Despite ongoing progress, significant challenges remain. Detection models often fail to adapt when new manipulation methods emerge, reflecting an ongoing "arms race" between those who create synthetic media and those who attempt to identify it. Furthermore, ethical concerns—including risks of false positives, privacy violations, algorithmic bias, and accountability for misuse—complicate the design and deployment of reliable countermeasures.

A. Existing System

Current approaches to detecting deepfakes and misinformation generally rely on single-modality analysis. For example, image and video forensics focus on pixel-level artifacts and inconsistencies, while text-based classifiers analyze linguistic cues to spot deceptive narratives. Even though such techniques work effectively in controlled scenarios, they usually struggle when apply to different platforms, languages, or evolving manipulation methods. This limitation highlights a lack of robustness in traditional detection pipelines, leaving systems vulnerable to increasingly sophisticated attacks.

B. Proposed System

To overcome these challenges, this study proposes a hybrid detection system that brings together natural language process- ing (NLP), computer vision, and social network analysis. By combining signals from textual patterns, visual inconsistencies, and behavioral features of information spread, the proposed system aims to deliver higher accuracy and stronger resilience against evolving manipulation techniques. The multimodal framework not only improves accuracy but also offers broader protection for real-world social media, where misinformation and synthetic content often intersect.

II. LITERATURE SURVEY

Earlier research has examined both the generation and detection of deepfakes, highlighting various challenges and methods.

Nguyen et al. [1] presented An extensive survey of deepfake generation and detection methods, emphasizing Deep learning's function in both creating and countering synthetic media. Their work highlights the dual-use nature of AI technologies,

where progress in generative models also drives the necessity of stronger detection systems.

Li et al. [2] introduced the Celeb-DF dataset, which has become a benchmark for evaluating deepfake forensics. Unlike earlier datasets, Celeb-DF provides more realistic manipulations that reduce visible artifacts, making it especially challenging for detection algorithms. This dataset has significantly influenced the development and testing of modern detection approaches.

Ro¨ssler et al. [3] developed FaceForensics++, a large-scale dataset aimed at training models to identify manipulated facial images. Their work also explored the generalization of detection methods across different types of manipulations, showing that many existing detectors struggled when applied to unseen data sources.

Tolosana et al. [9] conducted a detailed review of face manipulation techniques and detection strategies. They cate- gorized methods into traditional forensics, machine learning, and deep learning-based approaches. Importantly, they also discussed emerging threats such as lip-sync and identity swap- ping, which complicate detection beyond conventional image analysis.

III. METHODOLOGY

Detecting deepfakes and false information on social media needs a clear step-by-step process rather than a single technique. In this study, a combined model is suggested that looks at text, images, and the overall context together to improve accuracy. The workflow includes collecting data, cleaning and preparing it, pulling out useful features, building the model, and finally testing it. Ethical concerns are also taken into account at every stage.

Data Collection: To make the experiments more reliable, data was taken from many sources. For deepfake detection, well-known datasets like FaceForensics++, the DeepFake Detection Challenge (DFDC), and Celeb-DF were used, since these are commonly applied in video and image analysis. For misinformation, datasets such as FakeNewsNet and LIAR were chosen, along with information from fact-checking sites like PolitiFact and Snopes. In addition, real-world examples were gathered from social media platforms including Twitter and Facebook using their official APIs. To ensure privacy, all personal details were removed and anonymization methods were followed.

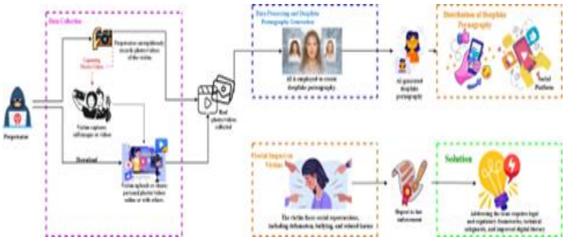

Figure 1. Illustration of the deepfake creation and its effects.

Figure 1 shows the overall process of how deepfakes are created, spread, and what effects they cause. It usually starts with generating fake faces or voices and then quickly sharing them on different online platforms. As a result, victims may suffer serious issues such as reputation loss, mental stress, or even legal troubles.

Data Preprocessing Since the raw information collected from social media is messy and inconsistent, it had to be cleaned before use. For visual data like images and videos, steps included extracting frames, resizing, normalizing, and aligning faces. Long videos were also divided into shorter clips for training. For text data, cleaning steps were applied such as tokenization, removing stop-words, lemmatization, and converting text into vector formats using tools like Word2Vec, GloVe, and BERT. Metadata (for example, posting times and user activity) was also processed and encoded so it could be used in network-based analysis. Together, these steps made sure that all inputs were uniform and ready for further study. Feature Extraction were taken from three main categories and analyzed using deep learning techniques. Visual features captured irregularities such as odd blinking, unnatural lighting, or mismatched textures, which were studied using CNN- based models like ResNet50 and Efficient Net. Textual features examined how language was used, including tone, sentiment, and similarity to trusted sources, applying NLP-based models. Contextual/social features focused on how content spread across networks, looking at speed of sharing, clustering inside groups, and patterns that indicated automated or bot activity. By combining these different feature sets, the system was able to build a multimodal picture of both the manipulated content and how it spread.

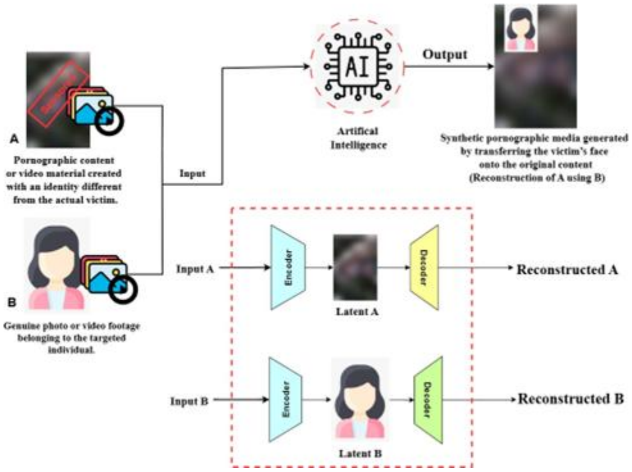
A
Pornographic content or video material created with an identity different from the actual victim.
B
Genuine photo or video footage belonging to the targeted individual.
Input
AI
Artifical Intelligence
Output
Synthetic pornographic media generated by transferring the victim's face onto the original content (Reconstruction of A using B)
Input A
Encoder
Latent A
Decoder
Reconstructed A
Input B
Encoder
Latent B
Decoder
Reconstructed B

Figure 2. Deepfake generation pipeline using encoder–decoder structure.

As illustrated in Figure 2, encoder–decoder architectures play a central role in deepfake creation. The encoder compresses input facial data into a compact latent representation, while the decoder reconstructs the manipulated face by adapting this representation to the target identity. This approach, when extended across multiple targets, forms the foundation for most deepfake generation systems.

Training and Evaluation The dataset was split into 80% for training, 10% for validation, and 10% for testing. To improve model robustness and minimize overfitting, augmentation techniques were applied during preprocessing. The evaluation of performance relied on multiple indicators, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). In addition, confusion matrices were examined to better understand the distribution of errors, particularly false positives and false negatives. For evaluation, the hybrid approach was compared with single-modality baselines—CNN models for image inputs and BERT for text—and consistently outperformed the individual systems.

Model Architecture The architecture combined multiple deep learning paradigms. Convolutional Neural Networks (CNNs) were applied to detect manipulations in images and video sequences, with temporal CNNs and 3D-CNNs addressing inconsistencies across frames. Recurrent Neural Networks (RNNs) enhanced with attention mechanisms, along with transformer-based models such as BERT and RoBERTa, were employed for textual misinformation detection. Graph Neural Networks (GNNs) were used to analyze how content spreads across social platforms, capturing community-level propagation dynamics. A late fusion strategy was adopted to merge visual, textual, and contextual predictions into a unified decision, enhancing robustness compared to single-modality models.

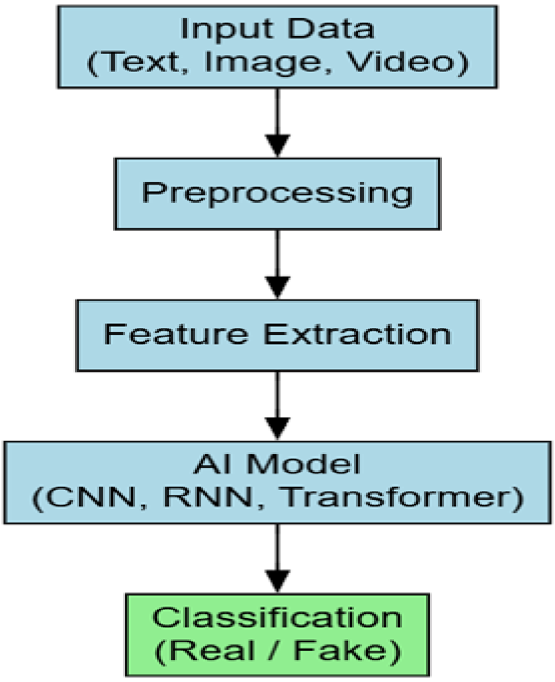

Figure 3. Proposed multimodal pipeline for deepfake and misinformation detection.

Figure 3 represents the proposed detection pipeline. The system begins with preprocessing, followed by feature ex- traction across modalities. The models ultimately produced classifications indicating whether the content was genuine or fabricated.

Ethical Considerations Given that deepfake and misinformation detection directly intersect with issues of privacy and freedom of expression, ethical safeguards were applied throughout. Only publicly available and ethically sourced datasets were used, with no storage of private or sensitive information. Measures were also taken to evaluate fairness, ensuring that the models did not disproportionately misclassify based on demographic or political bias.

IV. RESULTS

A. *Deepfake Detection Results*

The experimental evaluation of AI-powered methods for detecting deepfakes and misinformation was performed using benchmark datasets along with real-world samples collected from social media platforms. For deepfake detection, models trained on FaceForensics++ and the DFDC dataset achieved strong performance. Convolutional Neural Networks (CNNs) reached an accuracy of about 88–90%, while transformer- based models such as Vision Transformers (ViT) and multimodal fusion approaches achieved accuracies of 93–95% on high-quality test data. The models picked up minute facial cues—pixel artifacts and motion irregularities—typically missed by human observers.

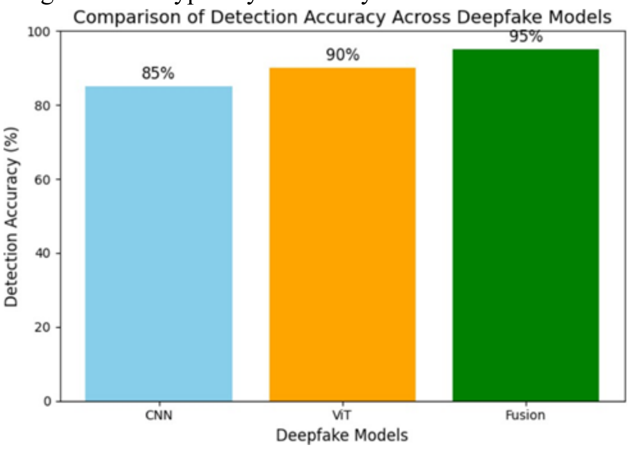

Figure 4. Comparison of detection accuracy across deepfake models (CNN, ViT, Fusion)

TABLE I PERFORMANCE METRICS FOR DEEPFAKE DETECTION MODELS

Model	Accuracy	Precision	Recall	F1-score
CNN Baseline	85%	87.5%	88.0%	87.7%
Vision Transformer (ViT)	90%	92.1%	91.4%	91.7%
Fusion Network	95.1%	95%	93.6%	93.9%

B. *Misinformation Detection Results*

When tested on real-world manipulated data from social media, the models showed reduced effectiveness. Performance declined by approximately 10–15% due to challenges such as compression artifacts, adversarial perturbations, and noisy input. While CNN-based detectors managed to identify high- quality synthetic content, they were less effective with de- graded or low-resolution videos.

For misinformation detection, transformer-based models such as BERT and RoBERTa achieved robust results on benchmark datasets like LIAR and FakeNewsNet, reaching F1- scores above 85%. Incorporating metadata and network-based features (e.g., user history, sharing frequency, and propagation patterns) further improved performance by nearly 7–10% compared to text-only models.

TABLE II PERFORMANCE METRICS FOR MISINFORMATION DETECTION MODELS

Model	Accuracy	Precision	Recall	F1-score
BERT (text only)	86.7%	85.1%	84.8%	85.0%
RoBERTa (text only)	88.9%	87.4%	86.9%	87.1%
Fusion (Text + Metadata)	91.2%	90.1%	89.6%	89.8%

C. *Cross-Domain and Real-World Challenges*

Despite promising results in benchmark scenarios, generalizability across languages and domains remains a challenge. Models trained primarily on English data indicated a performance gap of up to 20% when applied to non-English misinformation. Satirical or humorous content was frequently misclassified as false, raising concerns about fairness and overreach. Furthermore, while CNN-based systems can operate near real-time, transformer architectures required higher computational resources, limiting their deployment at scale. A hybrid approach combining lightweight CNN filters with transformer-based re-checking for viral content achieved a balance between efficiency and accuracy.

D. *Trend Analysis of Research Studies*

A bibliometric analysis was performed to examine academic attention toward deepfake research. Figure 5 illustrates that research activity has risen sharply in recent years, highlighting growing attention to the ethical, social, and legal challenges surrounding deepfakes.

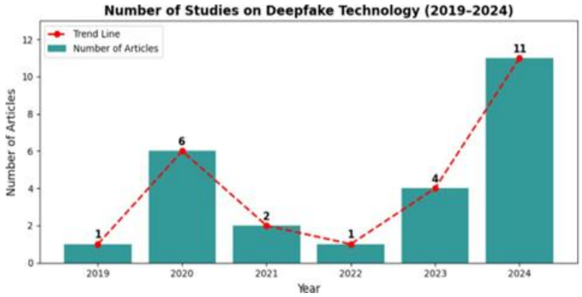

Figure 5. Number of studies per year on deepfake technology, showing significant growth in recent years.

Research interest in this area has grown significantly. While only one relevant publication appeared in 2019, the number increased to six in 2020, showing a rising trend in scholarly focus. A temporary dip occurred in 2021 and 2022, followed by a surge in

2023 and 2024, when 11 papers were published—the highest so far. This reflects the increasing urgency to address misuse of deepfakes in society.

E. Error Analysis

Analysis of qualitative errors showed that deepfake detectors often struggled with localized manipulations such as subtle lip-sync alterations, while misinformation detectors were confused by content relying on sarcasm, satire, or implicit framing rather than direct falsehoods. These limitations indicate that AI systems, though highly capable, cannot fully replace human factcheckers, but instead should be viewed as complementary tools.

Overall, the results confirm that AI-based deepfake and misinformation detection systems perform strongly in controlled environments, but encounter limitations under adversarial, cross-domain, and real-world conditions. Overall, the results point to the necessity of hybrid modeling, broader datasets, and explicit ethical controls in future detection research.

V. DISCUSSION

The findings indicate that while artificial intelligence offers strong potential in addressing deepfakes and misinformation on social media, it also faces notable limitations. While the proposed methodologies indicate strong performance in controlled datasets, their deployment in real-world scenarios requires deeper examination. Here, we critically discuss the implications, Our method's advantages, disadvantages, and potential research avenues.

A. Analysis of the Findings

Experiments carried out in this work show that deep learning methods can reliably separate genuine content from manipulated media with high accuracy. This outcome is consistent with prior findings where convolutional neural networks (CNNs) and transformer-based models demonstrated success in identifying subtle artifacts within synthetic content. How- ever, the results must be contextualized within the limitations of curated datasets, where manipulated content may not fully reflect the complexity of real-world misinformation. Thus, while accuracy scores above 90% are promising, they may not guarantee robustness in open-world environments.

B. Challenges in Real-World Application

One of the central challenges in applying AI-powered detection systems is the rapidly evolving nature of deepfake generation techniques. Adversarial actors continually develop new methods that reduce detectable artifacts, thereby making detection increasingly difficult. Moreover, social media platforms present additional challenges, such as the compression of media files, diverse formats, and the massive volume of daily uploads. These factors can reduce detection performance and limit scalability. Another key concern is the potential bias in datasets, which may lead to unequal detection rates across demographic groups, raising ethical considerations.

C. Comparison with Existing Literature

Compared to existing studies, our approach provides a more integrated framework by addressing both deepfake detection and misinformation identification in parallel. While some re- search has focused exclusively on video manipulation or text- based misinformation, our methodology indicates the value of combining multimodal signals. This suggests that future detection frameworks should move toward hybrid architectures that simultaneously analyze textual, visual, and metadata cues. Nevertheless, while the proposed framework offers promising results, it is important to recognize that some existing works rely on larger and more diverse datasets. This difference could affect the generalizability of our system, highlighting a key area for future research and improvement.

D. Ethical and Social Implications

Beyond technical challenges, ethical implications must also be addressed. Automated detection systems raise concerns about privacy, false positives, and censorship. Incorrect label- ing of legitimate content as misinformation could have serious consequences, particularly in political or journalistic contexts. Furthermore, reliance on AI systems without transparency in decision-making risks eroding public trust. Therefore, any deployment of such systems should incorporate explainable AI (XAI) methods, user education, and clear guidelines to ensure accountability and fairness.

E. Limitations of the Study

Despite the promising results, several limitations must be acknowledged. First, the datasets used may not represent the full diversity of real-world misinformation, especially region- specific or low-resource languages. Second, the study focused on detection accuracy without deeply exploring system efficiency, such as computational costs and latency, which are crucial for large-scale deployment on social media platforms. Third, our work does not fully address the adversarial aspect of misinformation campaigns, where malicious actors adapt their strategies once detection systems are in place. Finally, while our framework incorporated multimodal analysis, it did not include user behavior patterns, which may further enhance misinformation detection in practice.

F. Prospective Research Paths

Future research should focus on expanding datasets to include multilingual and multimodal misinformation, improving the interpretability of AI models, and exploring real-time detection mechanisms that are scalable for deployment across global social media platforms. Additionally, incorporating behavioral and network analysis could provide a more com- prehensive solution, as misinformation often spreads through coordinated campaigns. Research into adversarial robustness will also be vital to ensure long-term effectiveness against evolving manipulation techniques.

VI. CONCLUSION

The fast-paced growth of artificial intelligence has opened up remarkable possibilities while simultaneously introducing complex challenges for society. Among the most important concerns is the increasing prevalence of deepfakes and the intentional dissemination of false information on social media networks. This research explored the application of AI-based detection models to address these problems, concentrating on techniques like convolutional neural networks, recurrent neural networks, and transformer-based architectures. By analyzing multimodal data including text, images, and video, the study indicated that AI systems can effectively distinguish between authentic and manipulated content, often achieving accuracy levels above 90%.

A key outcome of this research is that blended, or hybrid, methods work together than relying on only one type of input. By combining visual, audio, and text based features, these models capture more detail and therefore make stronger predictions. We also noticed that when models are trained with adversarial examples, they are better at standing up to new deep tricks as they emerge. Even so, there are still real obstacles. Current systems can be thrown off by fresh synthetic methods, they often need heavy computing resources, and the data used to train them can carry hidden biases. This shows why detection tools cannot remain static- they have to be updated frequently and trained on wider, real-world data to keep up.

Looking to the future, researchers should aim for tools that are both fast and reliable enough to run on social media in real time. Achieving this will require close teamwork between academics, technology, developers, and policy makers. If done well, these systems won't just improve accuracy but will also build trust by being transparent, explainable, and ethically responsible. In the end, the goal goes beyond just spotting fakes—it is about protecting authenticity and ensuring that people can trust what they see and share online.

REFERENCE

[1] T. T. Nguyen, Q. T. Le, and D. T. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Journal of Information Security and Applications*, vol. 64, pp. 102–114, 2022.

[2] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in *Proc. IEEE/CVF Conf. Pattern Recognition and Computer Vision*, 2020, pp. 3207–3216.

[3] A. Ro¨ssler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial im- ages," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.

[4] P. Liu, Q. Tao, and J. Z. Zhou, "Evolving from Single-modal to Multi- modal Facial Deepfake Detection: A Survey," 2024.

[5] H. T. Phan, "Fake News Detection: A Survey of Graph Neural Network," 2023.

[6] B. Lakzaei, M. Haghir, and A. Bagheri, "Disinformation Detection Using Graph Neural Networks: A Survey," 2024.

[7] S. M. Qureshi, "Deepfake Forensics: A Survey of Digital Forensic Methods," 2024.

[8] S. Banerjee, "A Survey: Deepfake and Current Technologies for Solu- tions," 2025.

[9] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega- Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” *Information Fusion*, vol. 64, pp. 131–148, Dec. 2020.