Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Fake News and Deepfake Detection System

Namami Saxena
Department of AIML
Moradabad Institute of Technology
Moradabad, India
namamisaxena68@gmail.com

Onam Sinha
Department of AIML
Moradabad Institute of Technology
Moradabad, India
onamsinha01@gmail.com

Kavya Khatri
Department of AIML
Moradabad Institute of Technology
Moradabad, India
kavyakhatri306@gmail.com

Vandana
Department of AIML
Moradabad Institute of Technology
Moradabad, India
vandananainwal28@gmail.com

Hina Hashmi
Department of AIML
Moradabad Institute of Technology
Moradabad, India

Abstract -The rapid growth of social media and digital platforms has increased the spread of fake news and deepfake content, posing serious risks to online trust, social order, and information security. Fake news mainly involves false or misleading textual content, whereas deepfakes consist of artificially generated images or videos created to closely resemble real individuals. This review paper examines existing techniques used to detect both text-based misinformation and multimedia deepfakes. Detection approaches are broadly classified into machine learning methods for fake news analysis, deep learning techniques for image and video examination, and hybrid multimodal systems that integrate textual and visual features. Recent progress in natural language processing, convolutional neural networks, and transformer-based models is discussed along with commonly used evaluation measures and implementation strategies. A comparative study of current detection methods is presented by analyzing accuracy, computational requirements, interpretability, dataset limitations, and real-world applicability. The study observes that while individual detection models perform effectively in controlled environments, hybrid and multimodal approaches offer greater robustness against continuously evolving manipulation techniques. These integrated methods show strong potential for developing reliable, scalable, and accurate systems to safeguard the digital information ecosystem.

I. INTRODUCTION

Nowadays, many people prefer social media platforms and online news websites for searching and reading news instead of traditional newspapers. Although social media has become a powerful medium for information sharing, it has also negatively affected society by influencing major social and political events. After the 2016 U.S. presidential election, the problem of online misinformation gained significant statements and false claims can easily influence voters. Studies also show that false information spreads faster among people than genuine news and creates serious social consequences.

Several researchers have worked on identifying fake content using different techniques. Anuwat Chaiwongyen et al. analyzed timbre and shimmer features to distinguish between real and fake speech by examining multiple audio and shimmer components. Yang Hou et al. highlighted the impact of human circadian rhythm and proposed a statistical consistency attack to reduce differences in deepfake detection systems. Abu Qais et al. introduced a speech spoofing detection approach using Convolutional Neural Networks to differentiate between human speech and synthetic voices.

Other studies focused on improving the reliability of deepfake detection models. Muxin Pu et al. used metamorphic testing to evaluate the performance of the MesoInception-4 model. Chang-Sung Sung et al. proposed an audio-visual temporal synchronization framework to detect deepfakes, including unseen cases. Bo Zou et al. developed a contrastive pretraining framework that requires only a small amount of labeled data. Hefei Ling et al. presented a local-prediction method to supervise image regions, while Yun Huang et al. introduced DF-VLAD to aggregate multiple frames for improved detection performance.Deep Learning has recently emerged as an effective technology for fake news detection. Compared to traditional machine learning methods, deep learning approaches provide better accuracy in identifying false information. The availability of advanced programming frameworks has further encouraged the use of deep learning techniques. Consequently, many research studies published in recent years focus on deep learning-based methods for fake news detection

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

II. Literature Review

This SLR aimed to create a body of knowledge of deepfake detection techniques and to conduct a systematic review of the currently available literature regarding these techniques. The main objective was to undertake an SLR that analyses the effectiveness of deepfake detection techniques [Citation1]. Our specific objectives were to:

- Examine and analyse the current state of deepfake, providing an up-to-date overview of recent research work on deepfake detection.
- Analyse and measure the performance of various deepfake detection techniques using multiple metrics; and
- Identify and discuss significant advances, challenges, and future trends.

Deepfake detection targets synthetically generated or manipulated audio/video content (face swaps, reenactment, voice cloning). The problem is framed both as a forensic detection task (is this media manipulated?) and as a robustness/generalization challenge (detecting unseen generation methods).

A. Visual-Based Detection Approaches

Visual-based detection approaches focus on identifying spatial-level inconsistencies and visual artifacts present in manipulated images and video frames. Early research in this category relied on handcrafted forensic features such as abnormal eye blinking patterns, facial warping, unnatural skin textures, color mismatches, and inconsistencies in lighting and shadows. These cues were effective in detecting early deepfake generation methods but gradually became less reliable as synthesis techniques improved.

Recent studies have shifted toward deep learning-based visual analysis, particularly using Convolutional Neural Networks (CNNs). CNN-based models automatically learn hierarchical facial representations from raw pixel data, eliminating the need for manual feature engineering. Popular architectures such as VGGNet, ResNet, and XceptionNet have been widely adopted for frame-level deepfake detection. Among these, XceptionNet has demonstrated strong performance due to its depthwise separable convolutions, which efficiently capture subtle manipulation patterns. Despite their effectiveness, visual-based approaches often struggle with cross-dataset generalization, especially when tested on unseen or novel deepfake generation methods.

B. Audio-Based Detection Approaches

Audio-based deepfake detection approaches target manipulated or synthetically generated speech produced using voice cloning and text-to-speech technologies. Traditional audio forensic techniques focus on extracting handcrafted features such as Mel-Frequency Cepstral Coefficients (MFCCs), pitch variation, formant frequencies, and phase-based features. These methods aim to capture irregularities in speech signals that are difficult for generative models to replicate accurately.

With the advancement of deep learning, recent research has employed CNN and transformer-based architectures applied to spectrogram representations of audio signals. Spectrogram-based CNN models learn discriminative time–frequency patterns that differentiate real speech from synthetic audio. Transformer-based models further enhance detection by modeling long-range dependencies in speech signals. While audio-based approaches are effective in identifying voice deepfakes, their performance is often affected by background noise, compression artifacts, and variations in language and speaker characteristics.

C. Video Temporal Analysis Approaches

Video-based temporal analysis approaches aim to capture temporal inconsistencies across consecutive video frames that arise during deepfake generation. These inconsistencies may include unnatural facial movements, irregular lip synchronization, inconsistent head poses, and abrupt changes in facial expressions. Temporal modeling is particularly important because frame-based methods alone may fail to detect manipulations that appear visually realistic in individual frames.

To address this, researchers have proposed hybrid architectures combining CNNs with sequence models such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU). In these models, CNNs extract spatial features from individual frames, while LSTM or GRU layers analyze temporal dependencies across frame sequences. These approaches have shown improved robustness compared to static visual models; however, they require large-scale annotated video datasets and higher computational resources, making real-time deployment challenges.

D. Hybrid and Multimodal Detection Approaches

Hybrid and multimodal detection approaches integrate multiple modalities, including visual, audio, and temporal features, to enhance deepfake detection accuracy. These systems exploit cross-modal inconsistencies that often exist in manipulated media, such as mismatched lip movements and speech or inconsistent audio–visual synchronization. Feature fusion techniques are employed at different levels, including early fusion, late fusion, and attention-based fusion strategies.

Recent studies have also explored the incorporation of contextual and metadata information, such as social media propagation patterns and source credibility, using graph-based learning techniques. Multimodal approaches generally outperform unimodal systems due to their comprehensive analysis of content; however, they introduce increased system complexity, higher computational costs, and dependence on synchronized multimodal datasets.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

Ref No.	Authors	Methodology	Benefits	Limitations
1	Zhang, Zhao, and Li	Using a binary classifier trained by a CNN.	The achieved accuracy was 97%	Poor robustness
2	Lee et al.	false face detection pipeline that can identify fake face pictures.	Obtaining 93.99% accuracy	Robustness is not taken into account
3	Guo et al.	preprocessing module named AMTEN for face image forensics.	AMTENnet achieves an average accuracy of up to 98.52%	Detector's robustness is low
4	Guarnera et al.	expectation–maximization method trained to identify fingerprint	Obtaining a 93% accuracy rate	High delay
5	I.-J. Yu et al.	Use of MCNet to utilize multi-domain spatial, frequency, and compression domain characteristics.	High robustness	High complexity
6	J. Yang et	Using the image saliency to determine the texture depth and pixel difference between actual and fake facial images.	Detection accuracy is 0.9990	Poor robustness
7	Hsu et al.	Presenting an image detector comprised of an enhanced DenseNet backbone network and Siamese network architecture.	Achieving a modest level of precision and recall	High energy consumption
8	Güera & Delp	Advanced Video and Signal Based Surveillance	Provide Robustness	Expensive setup; complex system
9	Kohli and Gupta	Multimedia Tools and Applications	Works on low-power hardware	Lower accuracy than large models
10	Caldelli et al.	Pattern Recognition Letters	Captures long-term patterns	Requires GPU for training

TABLE I
SUMMARY OF RESEARCH STUDIES ON FAKE NEWS & DEEPFAKE DETECTION TECHNIQUES

III. METHODOLOGY

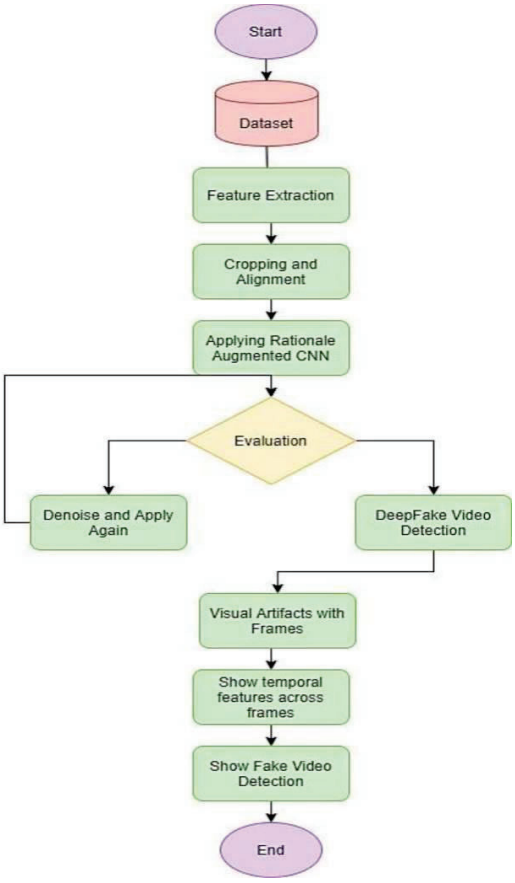

Fig. 1. Flowchart of Fake News & Deepfake Detection System

This study adopts an **experimental and comparative research design** to develop and evaluate an automated system for **fake news detection (text-based)** and **deepfake detection (multimedia-based)**. The methodology emphasizes robustness, generalization, and real-world applicability rather than benchmark overfitting. The system is designed as a **modular, multimodal pipeline**, enabling independent evaluation of textual and visual components as well as their combined performance.

IV. FUTURE DIRECTIONS

A. Critical Research Gaps and Limitations

Although artificial intelligence has made noticeable progress in detecting fake news and deepfake content, a close examination of existing research reveals several unresolved issues that limit the effectiveness of these systems in real-world environments.

1. **Adaptive and Continual Learning Approaches:** Future detection systems should be capable of learning continuously and adapting to new fake news patterns and emerging deepfake generation methods. Since misinformation strategies evolve rapidly on social media platforms, static models often become outdated and less effective over time.
2. **Cross-Domain and Cross-Platform Generalization:** There is a strong need for detection models that can perform consistently across different domains, platforms, and time periods. Research should focus on domain-independent feature learning and transfer learning techniques to improve robustness in real-world deployments, where data distributions frequently change.
3. **Support for Multilingual and Low-Resource Languages:** Most existing systems are designed primarily for high-resource languages. Future studies should address misinformation in regional and low-resource languages by developing multilingual and language-agnostic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

representations, as these languages are increasingly targeted in coordinated misinformation campaigns.

4. **Explainable and Transparent Detection Models:** Many deep learning-based systems function as black boxes, which limits trust and interpretability. Incorporating explainable AI techniques can help provide clear and understandable reasons behind detection decisions, improving transparency and user confidence.
5. **Efficient Multimodal Fusion Techniques:** Future research should explore flexible and lightweight multimodal fusion strategies that can dynamically adjust the importance of text, audio, and visual cues based on their availability and reliability, while also reducing computational overhead.
6. **Robustness Against Adversarial Manipulation:** Detection models must be designed to withstand adversarial attacks and intentional content manipulation. Techniques such as adversarial training, data augmentation, and invariant feature learning can help improve system stability under challenging real-world conditions.
7. **Early-Stage and Real-Time Detection:** Most current approaches detect fake news only after it has already spread widely. Future systems should aim to identify misinformation at early stages or in real time, allowing timely intervention before large-scale propagation occurs.
8. **Ethical, Privacy-Aware, and Regulation-Compliant Design:** Future detection frameworks should incorporate privacy-preserving learning methods and ethical considerations to ensure responsible deployment. Addressing regulatory and ethical challenges is essential for public acceptance and large-scale adoption.
9. **Standardized Evaluation and Benchmarking:** The absence of standardized datasets, metrics, and evaluation protocols makes comparison across studies difficult. Establishing common benchmarks would improve reproducibility and accelerate the transition of research models into practical systems.
10. **Incorporation of Social Context and Propagation Patterns:** Most existing methods focus mainly on content-level features such as text semantics or visual artifacts. Future research should also consider social context, user behavior, and information propagation dynamics to improve detection accuracy.

B. Future Research Directions:

Based on the identified challenges, future research should focus on developing practical, adaptive, and scalable detection systems capable of operating effectively in real-world environments. Emphasis should be placed on robustness, interpretability, ethical deployment, and real-time performance to address the growing impact of fake news and deepfake technologies.

V. CHALLENGES AND RESEARCH DIRECTIONS

Despite the large number of studies carried out on fake news detection, there is still considerable scope for further improvement and exploration. Although deep learning–based approaches generally achieve better accuracy than traditional methods, several challenges remain that limit their effectiveness and acceptance in real-world applications. Based on existing research, the following areas highlight important directions for future work.

1) The performance of fake news detection models is strongly influenced by the choice of features and classifiers. Many earlier studies did not give sufficient importance to this aspect. Long textual data often requires sequence-based models such as RNNs, yet limited research has focused on this requirement. Greater emphasis on feature engineering and classifier selection could lead to improved performance.
2) Most existing methods rely heavily on news content and headline features, while other factors such as user behavior, user profiles, and social network interactions remain underexplored. Features related to political or religious bias, along with lexical, syntactic, and statistical attributes, may enhance detection accuracy. Combining deep textual representations with these additional features could produce better results.
3) Research on news propagation patterns is limited in this domain. Network-based dissemination behavior has not been fully utilized for fake news identification. Studying how information spreads across social platforms may provide valuable insights for improving detection systems.
4) Many studies focus only on textual data, even though fake news is often created using a combination of manipulated text and images. Considering multiple data modalities is therefore necessary for more reliable detection.
5) Only a small number of studies have incorporated visual features such as images and videos. Expanding the use of visual data could improve the identification of manipulated or misleading content.
6) Detection models that learn from newly published online content in real time may enhance adaptability and accuracy. The use of transfer learning techniques with continuous data streams represents a promising direction for future research.
7) While CNN, LSTM, and ensemble-based approaches have achieved strong performance, models such as SeqGAN and Deep Belief Networks have received limited attention in this domain. Exploring these architectures may lead to improved detection capabilities.
8) Transformer-based models have largely replaced traditional RNN architectures in natural language processing tasks. Although BERT has been applied to fake news detection, the use of Generative Pre-trained Transformer models in this area remains limited. Fine-tuning such models for fake news detection may further improve results.
9) Ensemble learning approaches generally outperform single classifiers. Combining deep learning and machine learning models, where one model processes textual content and another handles auxiliary features, may lead to better detection performance.
10) The use of explainable AI techniques can help users understand why content is classified as fake or genuine. Providing clear explanations can increase trust and transparency in detection systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 04, ICTEM 2.0 (2026)

11) Although several datasets focusing on news content are publicly available, datasets covering diverse textual and hidden features are limited. Additional unexplored features may have a significant impact on detection performance, indicating the need for further investigation into richer and more varied feature sets.

VI. Conclusion

The rapid growth of deepfake video generation is creating serious technical as well as ethical challenges. With continuous improvements in machine learning techniques, especially generative models, deepfake content is becoming increasingly realistic and difficult to detect. This situation raises concerns about the misuse of such technology and highlights the importance of strengthening deepfake detection methods to protect digital media from manipulation.Although deep learning models have shown promising results in detecting deepfake videos and images, the quality of synthetic content is also improving at a fast pace. As a result, existing detection techniques require further enhancement to maintain accuracy. Another challenge lies in identifying suitable deep learning architectures, as there is still no clear understanding of the optimal model depth or structure needed for effective deepfake detection.Real-world conditions further complicate the performance of detection systems. Factors such as low-quality videos, compression effects, and rapidly evolving manipulation techniques often reduce the reliability of detection models that perform well in controlled environments. In many cases, deepfake generation methods advance faster than detection approaches, making it difficult to consistently identify manipulated content.Integrating deepfake detection mechanisms into social media platforms can play an important role in reducing the spread and impact of manipulated media. Such integration would allow platforms to respond more effectively to misinformation and minimize its negative influence on society. Continuous research is therefore essential to adapt detection systems to emerging threats.

In conclusion, future research should focus on the development of multimodal detection approaches and the establishment of standardized benchmarks for evaluation. Combining audio and visual information can provide more reliable detection, especially when manipulation in one modality is hidden by another. These efforts are crucial for ensuring the integrity, reliability, and security of digital content in complex and unpredictable real-world environments.

.

References

[1] H. Allcott and M. Gentzkow, ‘‘Social media and fake news in the 2016 election,’’ J. Econ. Perspect., vol. 31, no. 2, pp. 36–211, 2017.
[2] T. Rasool, W. H. Butt, A. Shaukat, and M. U. Akram, ‘‘Multi-label fake news detection using multi-layered supervised learning,’’ in Proc. 11th Int. Conf. Comput. Autom. Eng., 2019, pp. 73–77.
[3] X. Zhang and A. A. Ghorbani, ‘‘An overview of online fake news: Characterization, detection, and discussion,’’ Inf. Process. Manage., vol. 57, no. 2, Mar. 2020, Art. no. 102025. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306457318306794
[4] Abdullah-All-Tanvir, E. M. Mahir, S. Akhter, and M. R. Huq, ‘‘Detecting fake news using machine learning and deep learning algorithms,’’ in Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC), Jun. 2019, pp. 1–5.
[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, ‘‘Fake news detection on social media: A data mining perspective,’’ ACM SIGKDD Explorations Newslett., vol. 19, no. 1, pp. 22–36, 2017
[6] R. Oshikawa, J. Qian, and W. Y. Wang, ‘‘A survey on natural language processing for fake news detection,’’ 2018, arXiv:1811.00770.
[7] S. B. Parikh and P. K. Atrey, ‘‘Media-rich fake news detection: A survey,’’ in Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR), Apr. 2018, pp. 436–441..
[8] A. Habib, M. Z. Asghar, A. Khan, A. Habib, and A. Khan, ‘‘False information detection in online content and its role in decision making: A systematic literature review,’’ Social Netw. Anal. Mining, vol. 9, no. 1, pp. 1–20, Dec. 2019.Recent DL Works et al., “Transformer/CNN temporal modeling for driver fatigue,” 2020.
[9] Arefnezhad et al., “EEG + Bayesian filtering for continuous drowsiness estimation,” 2018.
[10] A. Bondielli and F. Marcelloni, ‘‘A survey on fake news and rumour detection techniques,’’ Inf. Sci., vol. 497, pp. 38–55, Sep. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025519304372
[11] P. Meel and D. K. Vishwakarma, ‘‘Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities,’’ Expert Syst. Appl., vol. 153, Sep. 2020, Art. no. 112986.
[12] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, ‘‘Combating fake news: A survey on identification and mitigation techniques,’’ ACM Trans. Intell. Syst. Technol., vol. 10, no. 3, pp. 1–42, May 2019.Florez et al., “CNN eye+mouth detection on Jetson Nano,” 2019.
[13] J. Ding, Y. Hu, and H. Chang, ‘‘BERT-based mental model, a better fake news detector,’’ in Proc. 6th Int. Conf. Comput. Artif. Intell., New York, NY, USA, Apr. 2020, pp. 396–400, doi: 10.1145/3404555.3404607.
[14] A. Giachanou, G. Zhang, and P. Rosso, ‘‘Multimodal multi-image fake news detection,’’ in Proc. IEEE 7th Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2020, pp. 647–654.
[15] D. Mangal and D. K. Sharma, ‘‘Fake news detection with integration of embedded text cues and image features,’’ in Proc. 8th Int. Conf. Rel., INFOCOM Technol. Optim., Trends Future Directions (ICRITO), Jun. 2020, pp. 68–72.
[16] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, ‘‘Exploiting multi-domain visual information for fake news detection,’’ in Proc. IEEE Int. Conf. Data Mining (ICDM), Nov. 2019, pp. 518–527
S. M. Padnekar, G. S. Kumar, and P. Deepak, ‘‘BiLSTM-autoencoder architecture for stance prediction,’’ in Proc. Int. Conf. Data Sci. Eng. (ICDSE),Dec.2020,pp.1–5.