

Research Paper

MULTIMODAL FAKE NEWS DETECTION USING DEEPPFAKE

GUIDE

Name : Mrs. J. BHANU SRI juluri.bhanusri@gmail.com

Assistant Professor

G. ABHINAYA - galipelliabhinaya04@gmail.com

B. AKSHAYA - akshaya.yadav170@gmail.com

B. GOUTHAMI - bangutapugouthami@gmail.com

A. SWETHA swetha.adepu05@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING VIGNAN'S INSTITUTE OF MANAGEMENT AND TECHNOLOGY FOR WOMEN

(An Autonomous Institution)

(Affiliated to Jawaharlal Nehru Technological University Hyderabad, Accredited by NBA, NAAC with A+)

Kondapur(Village), Ghatkesar (Mandal), Medchal (Dist.)

Telangana-501301

(2025-2026)

ABSTRACT

The rapid advancement of artificial intelligence has led to the widespread creation of synthetic media, including deepfake videos, AI-generated images, and cloned audio, posing serious threats to information authenticity and digital trust. This project presents **FakeDetect**, a robust multimodal deep learning framework designed to identify fake news and manipulated content across multiple data modalities, including text, images, audio, and video.

The proposed system integrates state-of-the-art models such as transformer-based architectures for text analysis, EfficientNet V2 for image verification, Wav2Vec2 for audio processing, and Xception networks for video deepfake detection. Each modality is independently analyzed to extract meaningful features and generate classification outputs with confidence scores. These outputs are then combined using a multimodal fusion mechanism to improve detection accuracy and reliability.

The system incorporates advanced preprocessing techniques, probabilistic

classification models, and temporal aggregation methods to ensure effective detection even in complex real-world scenarios. Additionally, FakeDetect provides explainable outputs such as heatmaps and frame-level analysis, enhancing transparency and user trust. Experimental results demonstrate that the system achieves high accuracy (approximately 92%) across diverse datasets, making it suitable for applications in media verification, cybersecurity, and digital forensics.

Keywords

Deepfake Detection, Multimodal Learning, Fake News Detection, Synthetic Media Forensics, Artificial Intelligence, Deep Learning, EfficientNet, Wav2Vec2, Xception Network, Media Authentication, Audio-Visual Analysis, Feature Fusion, Neural Networks, Content Verification, Misinformation Detection

I. INTRODUCTION

The rapid evolution of **Artificial Intelligence (AI)** and deep learning technologies has significantly transformed the digital content landscape. Modern generative models such as Generative Adversarial Networks (GANs), transformer-based architectures, and diffusion models have enabled the creation

of highly realistic synthetic media, including images, videos, and audio. While these technologies have beneficial applications in entertainment, education, and media production, they have also introduced serious challenges related to misinformation and digital content authenticity.

One of the most concerning outcomes of this advancement is the rise of **deepfakes**, which are artificially generated or manipulated media designed to appear authentic. Deepfake videos can convincingly alter facial expressions and lip movements, while AI-generated audio can mimic human voices with high accuracy. Similarly, manipulated images and fabricated textual content are increasingly used in misinformation campaigns. These developments have made it difficult for humans to distinguish between real and fake content, leading to a growing threat to trust in digital information.

The impact of fake news and deepfake content is widespread and affects multiple domains. In the political domain, manipulated media can influence public opinion and disrupt democratic processes. In the financial sector, fake announcements or synthetic audio of executives can lead to market manipulation. In cybersecurity, voice cloning enables sophisticated social

engineering attacks, while in personal contexts, deepfake content can cause reputational damage and psychological harm. As a result, ensuring the authenticity and integrity of digital media has become a critical challenge in today's information-driven society.

Existing fake news detection systems primarily rely on **single-modality (unimodal)** approaches, such as text-based classification using Natural Language Processing (NLP) or image-based detection using convolutional neural networks. Although some systems incorporate multiple modalities, they often analyze limited combinations such as text and image or audio and video. These approaches fail to capture complex relationships between different media types, resulting in reduced accuracy and limited real-world applicability. Additionally, many existing systems lack scalability, real-time processing capability, and explainability, making them less effective in practical deployment scenarios.

To address these limitations, this project proposes a **multimodal fake news detection system**, called *FakeDetect*, which integrates analysis of text, images, audio, and video within a unified framework. The system leverages advanced deep learning models such as transformer-

based architectures for textual understanding, EfficientNet V2 for image analysis, Wav2Vec2 for audio processing, and Xception networks for video deepfake detection. By combining outputs from multiple modalities using a fusion mechanism, the system improves detection accuracy and robustness.

II. LITERATURE REVIEW

The rapid increase in fake news and deepfake content has driven extensive research in the field of **multimedia fake news detection**, with approaches evolving from traditional machine learning techniques to advanced deep learning and multimodal frameworks. Early studies primarily focused on **text-based fake news detection**, where Natural Language Processing (NLP) techniques such as TF-IDF, Support Vector Machines (SVM), and later transformer-based models like BERT were used to classify news articles as real or fake. While these approaches achieved reasonable performance, they were limited in handling multimedia content and failed to detect manipulations in images, audio, and video.

With the growing availability of multimedia data, researchers began exploring **image-based and video-based detection methods**. Convolutional Neural

Networks (CNNs) have been widely used to identify visual inconsistencies, artifacts, and unnatural patterns in manipulated images. Similarly, deep learning models such as Xception and ResNet have shown strong performance in detecting deepfake videos by analyzing facial features and spatial distortions across frames. However, these approaches are restricted to visual data and cannot capture inconsistencies present in other modalities like audio or text.

To overcome the limitations of unimodal systems, recent research has shifted towards **multimodal fake news detection**, which integrates multiple data sources such as text, images, and audio. One notable approach is the **Heterogeneous Fusion Network (HFN)**, which combines textual, visual, and audio features using a fusion mechanism to improve classification accuracy. HFN is effective in detecting inconsistencies across modalities but is primarily designed for short-form video content and lacks flexibility for broader applications.

Another significant system is **DeepTruth AI**, which utilizes BERT for text analysis, CNNs for image processing, and Long Short-Term Memory (LSTM) networks for audio analysis. This system enhances detection performance by incorporating

cross-modal consistency checks such as lip-sync analysis. Despite its advantages, it suffers from high computational complexity and limited scalability, making it less suitable for real-time deployment.

Recent advancements have introduced more sophisticated multimodal models such as **MAGIC (Multimodal Adaptive Graph Neural Network)** and **AMPLE (Emotion-Aware Multimodal Learning)**. The MAGIC model leverages graph neural networks to capture complex relationships between text and image data, achieving high accuracy on benchmark datasets. However, it is computationally expensive and prone to overfitting. AMPLE incorporates emotional and contextual features to improve detection but is limited to text and image modalities and may introduce bias due to emotional dependency.

In the domain of deepfake video detection, **AV-FakeNet** focuses on audio-visual consistency by analyzing synchronization between speech and lip movements. This approach is effective for identifying manipulated videos but ignores textual and image-based information, making it incomplete for comprehensive fake news detection. Additionally, **cross-modal large model fusion techniques** have recently emerged, combining large language models

with vision-based systems to enhance semantic understanding and reasoning across modalities. Although these approaches improve detection performance, they require large-scale datasets and are not suitable for real-time applications.

A recent **multimodal survey (2025)** highlights that integrating multiple modalities—text, image, audio, and video—provides the highest accuracy and robustness in fake news detection. However, it also identifies key challenges, including the lack of comprehensive datasets, difficulty in designing effective fusion strategies, and high computational requirements.

III. METHODOLOGY

1. Data Input and Acquisition

The system accepts multiple types of user inputs such as textual news content, images, audio clips, and video files. This multimodal input capability enables the system to handle real-world data from social media, news platforms, and digital sources. Each type of input is processed independently to ensure accurate analysis.

2. Data Preprocessing

Before analysis, the input data undergoes preprocessing to ensure consistency and improve model performance:

- **Text:** Tokenization, stop-word removal, and normalization
- **Image:** Resizing, normalization, and noise reduction
- **Audio:** Noise filtering, resampling, and waveform normalization
- **Video:** Frame extraction and resizing

3. Feature Extraction

The system employs specialized deep learning models to extract meaningful features from each modality:

- **Text Analysis:** Transformer-based models (e.g., BERT) capture contextual and semantic relationships
- **Image Analysis:** EfficientNet V2 extracts spatial features and detects visual artifacts
- **Audio Analysis:** Wav2Vec2 learns speech representations to identify synthetic or cloned voices
- **Video Analysis:** Xception network analyzes frame-level features to detect deepfake manipulations

4. Modality-wise Classification

Each modality is independently classified as **Real** or **Fake** using deep learning classifiers.

- The models generate probability scores using sigmoid or softmax functions
- Confidence scores indicate the likelihood of content being manipulated

5. Multimodal Fusion Mechanism

The outputs from all modalities are combined using a **fusion strategy**, such as weighted averaging or probabilistic fusion.

- Integrates complementary information from text, image, audio, and video
- Reduces false predictions and improves overall accuracy
- Captures cross-modal inconsistencies

6. Temporal Analysis for Video

For video inputs, frame-level predictions are aggregated using temporal techniques:

- Maximum aggregation
- Mean aggregation
- Weighted averaging

7. Output Generation and Visualization

The final output is presented to the user with:

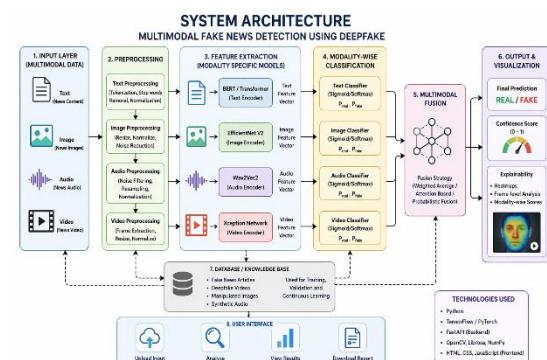
- **Prediction:** Real or Fake
- **Confidence Score:** Probability of classification
- **Explainability:** Heatmaps and frame-level analysis

8. System Implementation

The system is implemented using:

- **Backend:** Python with FastAPI for API handling
- **Deep Learning Frameworks:** TensorFlow and PyTorch
- **Libraries:** OpenCV, Librosa, NumPy
- **Frontend:** HTML, CSS, JavaScript

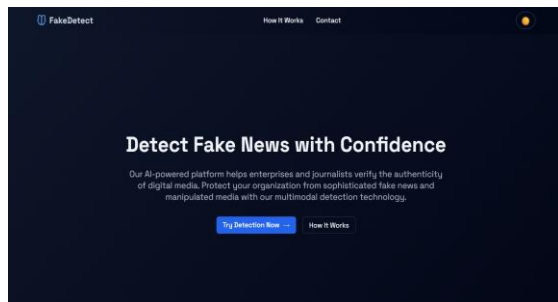
IV. SYSTEM ARCHITECTURE



V. RESULTS & DISCUSSION

The proposed **FakeDetect** system was evaluated using multiple benchmark datasets covering text, image, audio, and video modalities to assess its effectiveness

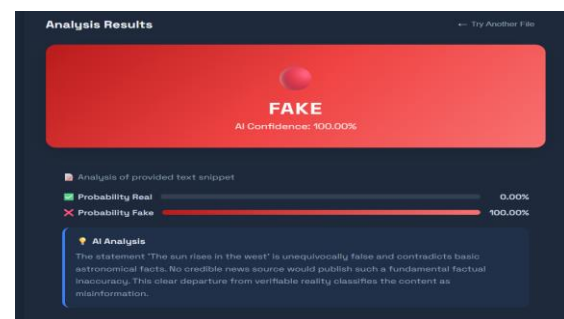
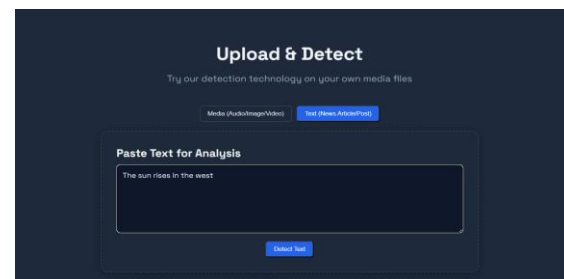
in detecting fake news and deepfake content. The experimental results demonstrate that the system achieves a **high overall accuracy of approximately 92%**, indicating strong performance across diverse types of multimedia data.



The evaluation was carried out using standard performance metrics such as **accuracy, precision, recall, F1-score, and AUC (Area Under Curve)**. The results show that the multimodal approach consistently outperforms traditional unimodal systems. This improvement is primarily due to the system's ability to capture complementary information from multiple sources, reducing the chances of misclassification. For instance, when textual content appears genuine but is accompanied by manipulated images or synthetic audio, the system successfully identifies inconsistencies through cross-modal analysis.

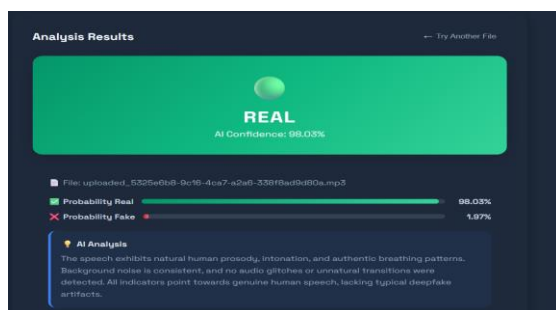
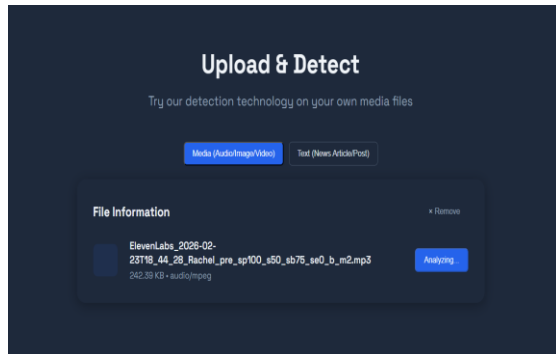
In the case of **image detection**, the EfficientNet V2 model effectively identifies visual artifacts, unnatural textures, and inconsistencies in facial

features, leading to high classification accuracy. For **audio analysis**, the Wav2Vec2 model demonstrates strong capability in detecting synthetic or cloned speech by analyzing speech patterns and frequency characteristics. Similarly, the **video detection module**, based on the Xception architecture, performs reliable frame-level analysis and successfully detects deepfake manipulations using temporal aggregation methods.



The **multimodal fusion mechanism** plays a crucial role in enhancing system performance. By combining predictions from all modalities using a weighted or probabilistic approach, the system achieves better generalization and robustness compared to single-modality models. This fusion strategy significantly reduces false positives and false negatives, especially in

complex real-world scenarios where manipulation may occur in only one modality.



Another important aspect of the system is its **explainability**. The inclusion of heatmaps and frame-level visualizations provides insights into which regions of an image or video are identified as manipulated. This improves user trust and makes the system more suitable for practical applications such as media verification and digital forensics.

VI. CONCLUSION

In this project, a comprehensive **multimodal fake news and deepfake detection system**, named *FakeDetect*, has been successfully designed and

implemented to address the growing challenges of misinformation and synthetic media. The system integrates advanced deep learning models to analyze multiple data modalities, including text, images, audio, and video, within a unified framework. By leveraging models such as transformer-based architectures, EfficientNet V2, Wav2Vec2, and Xception, the system is capable of identifying subtle inconsistencies and manipulation patterns across different types of media.

The incorporation of a **multimodal fusion mechanism** significantly enhances detection accuracy and reliability compared to traditional unimodal approaches. Experimental results demonstrate that the system achieves high performance (around 92% accuracy), making it suitable for real-world applications such as media verification, cybersecurity, and digital forensics. Additionally, the inclusion of explainable outputs, such as heatmaps and frame-level analysis, improves transparency and user trust in the system's predictions.

Despite its strong performance, the system faces certain limitations, including high computational requirements and challenges in processing low-quality or highly compressed data. These limitations

highlight the need for further optimization and scalability improvements.

REFERENCES

- [1] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014, pp. 2672–2680.
- [2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” in *Proc. International Conference on Learning Representations (ICLR)*, 2018.
- [3] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.
- [4] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- [5] A. Vaswani *et al.*, “Attention Is All You Need,” in *Proc. NeurIPS*, 2017, pp. 5998–6008.
- [6] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proc. ICML*, 2019, pp. 6105–6114.
- [7] A. Gulrajani *et al.*, “Improved Training of Wasserstein GANs,” in *Proc. NeurIPS*, 2017, pp. 5767–5777.
- [8] A. Rössler *et al.*, “FaceForensics++: Learning to Detect Manipulated Facial Images,” in *Proc. IEEE/CVF ICCV*, 2019, pp. 1–11.
- [9] B. Dolhansky *et al.*, “The DeepFake Detection Challenge (DFDC) Dataset,” *arXiv preprint arXiv:2006.07397*, 2020.
- [10] H. Li *et al.*, “Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” in *Proc. IEEE/CVF CVPR*, 2020, pp. 3207–3216.