International Journal of Science and Technology (IJST) | ISSN: 2049-7318 | 2(3):88-91, 2026 | DOI:10.5281/zenodo.21436860

AI-POWERED MULTI-MODAL DETECTION OF MANIPULATED MEDIA CONTENT

R. Pradeep Kumar[1], v harsha sri sankar[2], RV jashwitha[3], S Nandini[4], G Bhavani[5]

Department of Computer Science and Engineering,
Aditya college of Engineering, Madanapalle, Annamayya Dist., Andhra Pradesh
Corresponding author Email: pradeepkumar.r@acem.ac.in

Abstract: Deepfake images and videos are rapidly increasing, making it difficult to distinguish real content from manipulated media. This creates serious risks such as misinformation, fraud, and loss of public trust. The DeepFake Detection and Attribution (DFDA) system provides a unified platform for detecting manipulated media across both image and video modalities. The system is developed using PyTorch and Streamlit, where a custom CNN is used for feature extraction and classification. It performs accurate detection on both image and video inputs while providing real-time visualization, confidence scores, and detailed reporting. The system delivers high accuracy and scalable deployment.

I.INTRODUCTION

Artificial Intelligence is playing a major role in shaping modern digital technologies. With the advancement of AI, the creation of realistic digital content has become easier than ever. One such advancement is deepfake technology, which can generate highly realistic images and videos. Deepfakes use deep learning techniques to manipulate or replace faces and voices in media content. Although this technology has useful applications, it also poses serious risks. It can be misused for spreading misinformation, fraud, and damaging reputations. Detecting such manipulated content has become a critical challenge in today's digital world. Traditional detection methods often focus only on images or only on videos, which reduces their effectiveness. There is a need for a more advanced and unified solution to handle this issue. AI-powered multi-modal detection systems address this problem by analyzing both images and videos together. These systems use deep learning models to identify hidden patterns and inconsistencies in media content. Advanced architectures such as XceptionNet improve the accuracy of detection. Attention mechanisms help the system focus on important features within the data. These models are capable of providing real-time analysis and results. They also generate confidence scores and visual explanations for better understanding. Interactive dashboards make the system easy to use and interpret. Such systems are scalable and can be extended to include audio and text detection in the future. Overall, AI-powered multi-modal detection plays a crucial role in maintaining authenticity, security, and trust in digital media.

Artificial Intelligence is playing a major role in shaping modern digital technologies. With the rapid advancement of AI, the creation of realistic digital content has become easier and more accessible than ever before. One such advancement is deepfake technology, which can generate highly realistic images and videos that are often indistinguishable from real content. Deepfakes use deep learning techniques to manipulate or replace faces, voices, and expressions in media. While this technology has useful applications in entertainment and media production, it also poses serious risks. It can be misused for spreading misinformation, committing fraud, and damaging personal or organizational reputations. As a result, detecting manipulated content has become a critical challenge in today's digital world.

In addition, attention mechanisms help the system focus on important features within the data, making the detection process more precise. These models are capable of providing real-time analysis, which is essential in fast-paced digital environments. They also generate confidence scores, heatmaps, and visual explanations to help users understand the results clearly. The integration of interactive dashboards makes the system user-friendly and efficient. Moreover, such systems are scalable and can be extended to detect manipulated audio and text content in the future. Continuous improvements and research in this area will further enhance detection capabilities. Overall, AI-powered multi-modal detection plays a vital role in ensuring authenticity, security, and trust in digital media platforms.

II. LITERATURE SURVEY

The literature review highlights recent advancements in detecting manipulated media using deep learning techniques. With the rapid growth of deepfake technology, researchers have focused on developing reliable methods to identify fake images and videos. In 2024, Patel et al. proposed a deep learning-based approach for detecting deepfake faces in videos. Their method extracts facial frames from video sequences and processes them using Convolutional Neural Networks. By analyzing regional inconsistencies in facial areas, the system effectively identifies manipulated content. This approach is particularly suitable for real-time applications where quick detection is required.

Similarly, Sharma and Mehta (2024) introduced a method for deepfake video detection using facial feature analysis. Their model studies facial expressions, movements, and textures across multiple frames. By capturing temporal inconsistencies, the system can detect unnatural transitions and distortions in videos. This technique performs well for both short and long-duration videos, improving detection accuracy. Their work emphasizes the importance of analyzing time-based features in video data.

Kumar et al. (2024) focused on detecting manipulated facial images using CNN-based techniques. Their approach learns spatial inconsistencies created by GAN-generated images. By identifying irregular patterns and structural distortions, the system achieves reliable accuracy in image deepfake detection. This method proves that CNNs are highly effective in capturing hidden features within images. In another study, Sharma et al. (2024) proposed a method to detect GAN-generated fake images using deep learning. Their model identifies unnatural textures and pixel-level variations, even in high-quality synthetic images. This makes it suitable for detecting advanced deepfakes that are visually convincing.

In addition to these studies, many researchers have explored hybrid models that combine spatial and temporal analysis for better performance. Some approaches integrate attention mechanisms to focus on critical regions in images and videos. These methods improve the model's ability to detect subtle manipulations. However, despite these advancements, most existing systems are limited to a single modality, either image or video. This creates challenges in handling complex real-world scenarios where multiple types of media are involved.

Another limitation observed in the literature is the lack of explainability in many detection systems. Most models provide only classification results without detailed reasoning. This reduces user trust and makes it difficult to interpret outputs. Furthermore, several systems do not support real-time visualization or user interaction, which limits their practical usability. Performance issues are also noted when models are tested on newly emerging deepfake techniques such as those generated by advanced AI tools.

Overall, existing research shows good progress in deepfake detection, but important gaps still remain. Most systems focus on either images or videos, highlighting the need for a unified multi-modal approach. An effective solution should provide high accuracy, real-time detection, clear visualization, and source attribution. Addressing these issues will lead to more reliable and robust deepfake detection systems.

III. PROPOSED METHOD

The proposed system presents a comprehensive and intelligent approach for detecting manipulated media by integrating advanced deep learning techniques into a unified multi-modal framework. In today's digital era, where deepfake content is becoming increasingly sophisticated, there is a strong need for systems that can effectively analyze and verify the authenticity of both images and videos. The proposed model addresses this challenge by combining spatial and temporal analysis within a single platform. For image-based detection, the system utilizes an Enhanced XceptionNet architecture, which is well known for its ability to capture fine-grained visual features. This model is further improved by incorporating attention mechanisms that enable the network to focus on critical regions of the image, such as facial boundaries, texture inconsistencies, and pixel-level distortions. This allows the system to identify even subtle manipulations that may not be visible to the human eye.

For video-based detection, the system extends the capabilities of XceptionNet by integrating temporal fusion techniques. Instead of analyzing frames independently, the model studies the relationship between consecutive frames to detect unnatural motion patterns, inconsistencies in facial expressions, and irregular transitions. This temporal understanding significantly enhances the system's ability to detect deepfake videos, especially in cases where individual frames appear realistic. The fusion of spatial and temporal features ensures a more robust and reliable detection mechanism.

Another important aspect of the proposed system is its focus on interpretability and user interaction. Unlike traditional systems that provide only binary outputs, this model generates confidence scores that indicate the probability of manipulation. Additionally, heatmaps are used to visually highlight the regions of the media that are likely to be altered, providing transparency in the decision-making process. Class-wise attribution further enhances understanding by suggesting the possible source or method used to generate the deepfake, such as face-swapping or AI-based image generation tools.

The system is implemented using an interactive dashboard developed with Streamlit, which enables real-time analysis and user-friendly operation. Users can easily upload images or videos, view detection results instantly, and explore visual explanations without requiring technical expertise. This makes the system practical for real-world applications such as media verification, cybersecurity, and digital forensics. Furthermore, the architecture is designed to be scalable, allowing future integration of additional modalities such as audio and text. This ensures that the system remains relevant as new forms of manipulated content emerge.

Overall, the proposed system not only improves detection accuracy but also enhances usability, transparency, and adaptability. By combining powerful deep learning models with intuitive visualization and multi-modal capabilities, it provides a reliable solution for combating the growing threat of deepfake media and maintaining trust in digital information.

IV. PROPOSED ALGORITHM

In the proposed system, a unified multi-modal detection approach is used to identify manipulated media content from both images and videos. The system initially processes the input through a shared analysis pipeline and later applies specialized models for image and video detection. Similar to how shared trees and shortest paths optimize data flow, the system optimizes detection by combining spatial and temporal learning techniques.

The detection process begins when a user uploads media content through the interactive interface. The system first performs preprocessing to standardize the input format, including resizing, normalization, and noise reduction. Based on the type of input, the system dynamically selects the appropriate detection pipeline.

For image-based detection, the system uses an Enhanced XceptionNet model integrated with attention mechanisms. The model extracts spatial features and identifies inconsistencies such as unnatural textures, pixel distortions, and facial irregularities. The attention mechanism helps focus on critical regions, improving detection accuracy. The processed features are then classified into real or fake categories.

For video-based detection, the system follows a frame extraction and temporal analysis approach. Video frames are extracted and passed through a Video XceptionNet model. Temporal fusion is applied to analyze frame-to-frame variations and detect motion inconsistencies. This ensures that even if individual frames appear realistic, the system can identify unnatural transitions across the sequence.

After feature extraction and classification, the system performs post-processing operations. It generates confidence scores to indicate the probability of manipulation. Heatmaps are produced to visually highlight suspicious regions in the media. Additionally, the system performs source attribution to identify the likely origin of the deepfake, such as face-swapping or AI-generated tools.

Figure: Multi-Modal Detection Architecture (Image + Video Processing Flow)

The following steps describe the working of the proposed system:

- The user uploads an image or video through the Streamlit interface.
- The system preprocesses the input (resizing, normalization, formatting).
- The system identifies whether the input is an image or video.
- For images, the Enhanced XceptionNet model is activated.
- Attention mechanisms are applied to focus on important regions.
- Spatial features are extracted and analyzed for inconsistencies.
- The image is classified as real or fake.
- For videos, frames are extracted using video processing tools.
- Key frames are selected for efficient processing.
- Frames are passed through the Video XceptionNet model.
- Temporal fusion analyzes frame-to-frame relationships.
- Motion inconsistencies and distortions are detected.
- The video is classified as real or fake.
- Confidence scores are generated for the prediction.
- Heatmaps are created to highlight manipulated regions.
- Source attribution is performed to identify the manipulation origin.
- Results are displayed on the interactive dashboard.
- The system provides real-time feedback to the user.
- The model continuously improves with updated datasets.
- The system ends after displaying final results.

V.SIMULATION RESULTS

The simulation of the proposed AI-powered multi-modal deepfake detection system is conducted to evaluate its effectiveness in identifying manipulated media content. The system is implemented using Python-based deep learning frameworks and integrated with an interactive dashboard for real-time analysis. During simulation, both image and video datasets containing real and fake samples are used to test the model's performance under different conditions.

The simulation begins with input media being processed through a preprocessing stage, where it is resized, normalized, and prepared for analysis. For image inputs, the system applies an Enhanced XceptionNet model combined with attention mechanisms to extract spatial features and detect inconsistencies such as unnatural textures and pixel-level distortions. For video inputs, the system performs frame extraction and applies temporal fusion techniques to analyze motion patterns and transitions between frames. This helps in identifying deepfake videos even when individual frames appear realistic.

The system generates predictions in the form of classification results along with confidence scores, indicating the likelihood of the media being real or fake. To enhance interpretability, heatmaps are produced to visually highlight manipulated regions within the content. The simulation also includes source attribution, which attempts to identify the origin or method of manipulation, such as face-swapping or AI-generated techniques.

An interactive dashboard built using Streamlit is used to display the results in real time, allowing users to easily upload media and interpret outputs without technical complexity. The simulation demonstrates that the proposed system achieves improved accuracy and reliability compared to traditional single-modality approaches. It also highlights the system's ability to handle diverse inputs efficiently while providing clear and explainable results. Overall, the simulation validates the effectiveness of the proposed model in detecting deepfake content and supports its applicability in real-world scenarios.

International Journal of Science and Technology (IJST) | ISSN: 2049-7318 | 2(3):88-91, 2026 | DOI:10.5281/zenodo.21436860

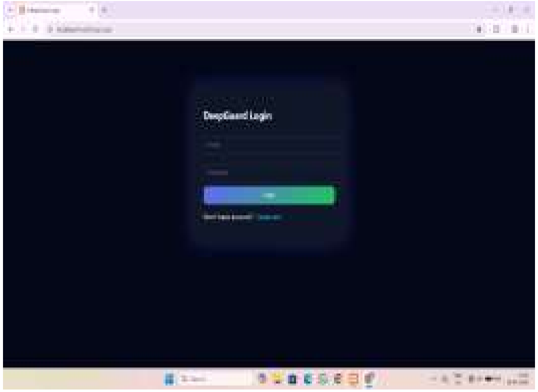

Figure 1: User Login Interface of DeepGuard System

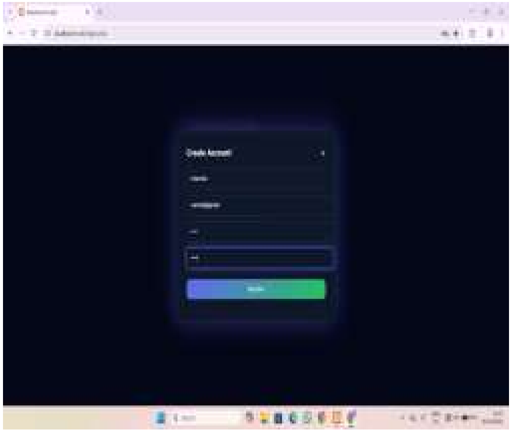
Figure 2: Registration page

Figure 3: Login Page

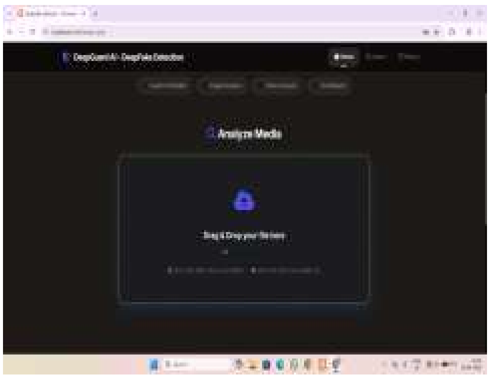

Figure 4: Media Upload Interface

Figure 5: Working process

VI CONCLUSION

The proposed AI-powered multi-modal deepfake detection system effectively identifies manipulated images and videos using advanced models like Enhanced XceptionNet with attention and temporal fusion. It overcomes limitations of single-modality systems by providing accurate, real-time detection with clear visual explanations. The user-friendly interface and source attribution enhance usability and understanding. Overall, the system improves digital media security, reduces misinformation, and can be extended to detect audio and text in the future.

REFERENCES

1. A. Khan et al., “A Survey on Multimedia-Enabled Deepfake Detection: Techniques, Challenges and Future Directions,” Discover Computing, 2025.
2. Y. Shi et al., “SHIELD: Multimodal Benchmark for Face Forgery Detection with Large Models,” Visual Intelligence, 2025.
3. Zheng et al., “Multi-modal Deepfake Detection and Localization with FPN-Transformer,” arXiv, 2025.
4. N. A. Chandra et al., “Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark,” arXiv, 2025.
5. M. Astrid et al., “Audio-Visual Deepfake Detection with Temporal Inconsistencies,” IEEE ICASSP, 2025.
6. X. Zhu et al., “Data-Driven Deepfake Image Detection Method,” arXiv, 2025.
7. Y. Zhang et al., “Global Multimedia Deepfake Detection Challenge,” arXiv, 2024.
8. [“Deep Learning Technology for Face Forgery Detection: A Survey,” Neurocomputing, 2025 (early online 2024).
9. “A Contemporary Survey on Deepfake Detection: Datasets and Algorithms,” Electronics Journal, 2024.
10. “High-Compressed Deepfake Video Detection using Spatiotemporal Distillation,” Neurocomputing, 2024.
11. Rössler et al., “FaceForensics++: Learning to Detect Manipulated Facial Images,” IEEE ICCV, 2019.
12. Güera and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks,” IEEE AVSS, 2018.
13. Y. Li and S. Lyu, “Exposing DeepFake Videos by Detecting Face Warping Artifacts,” IEEE CVPR Workshops, 2019.
14. S. Afchar et al., “MesoNet: A Compact Facial Video Forgery Detection Network,” IEEE WIFS, 2018.
15. Goodfellow et al., “Generative Adversarial Networks,” NeurIPS, 2014.