AI-Based Real Time Deep Fake Detection

Ms. S. Mahalakshmi[1], C. Akash[2], R. Vasanthakumar[3], R. Kalyanamoorthy[4], M. Sanjai[5]
[1]Assist prof, Dept of Artificial Intelligence and Data Science,
[2, 3, 4, 5] Dept of Artificial Intelligence and Data Science,
[1, 2, 3, 4, 5] Sir Issac Newton College of Engineering and Technology.

Abstract- *The rapid advancement of deep learning and generative models, particularly Generative Adversarial Networks (GANs), has enabled the creation of highly realistic synthetic media known as deepfakes. These manipulated media forms, including images, videos, audio, and text, pose significant threats to digital trust, cybersecurity, and social stability. Deepfakes are increasingly used for misinformation campaigns, identity theft, political manipulation, and financial fraud, making their detection a critical research challenge. This paper proposes a multimodal deepfake detection system that integrates advanced Artificial Intelligence techniques to analyze and classify content across multiple data modalities. The system employs BERT-based Natural Language Processing (NLP) for text analysis, Convolutional Neural Networks (CNNs) for image and audio classification, and Long Short-Term Memory (LSTM) networks for temporal video analysis. The proposed system is evaluated using benchmark datasets such as Celeb-DF, FaceForensics++, and ASVspoof, achieving high accuracy across all modalities. Furthermore, the system is implemented as a web-based platform that enables real-time detection of deepfake content. The results demonstrate that the proposed approach significantly improves detection performance and provides a scalable solution for combating misinformation in digital ecosystems.*

I. INTRODUCTION

In recent years, the proliferation of digital media has transformed the way information is created, shared, and consumed. Social media platforms, video-sharing services, and online communication channels have made it easier for individuals to access and distribute content globally. However, this accessibility has also led to the widespread dissemination of manipulated and misleading information.

Deepfake technology, powered by deep learning algorithms such as GANs and autoencoders, allows for the creation of highly realistic synthetic media that can mimic real individuals' faces, voices, and expressions. These technologies can generate convincing fake videos, alter speech, and fabricate textual content, making it increasingly difficult to distinguish between authentic and manipulated data. The consequences of deepfake technology are far-reaching. It can be used to spread fake news, manipulate public opinion, damage reputations, and even facilitate cybercrimes such as identity fraud.

Traditional detection methods, which rely on manual verification or rule-based systems, are inadequate in handling the scale and complexity of modern deepfakes. To address these challenges, there is a need for an automated, intelligent, and scalable detection system that can analyze multiple forms of media simultaneously. This paper presents a multimodal deepfake detection framework that integrates state-of-the-art AI techniques to provide accurate and real-time detection.

II. LITERATURE REVIEW

Deepfake detection has become an active area of research in artificial intelligence and cybersecurity. Various techniques have been proposed to address this problem, each focusing on different aspects of media analysis. In image-based detection, Convolutional Neural Networks (CNNs) have been widely used to identify visual artifacts and inconsistencies introduced during the generation process. Similarly, video-based detection methods utilize Recurrent Neural Networks (RNNs) and LSTM models to analyze temporal patterns and detect anomalies across frames.

For text-based misinformation detection, transformer-based models such as BERT have demonstrated significant improvements in understanding contextual relationships and semantic meaning. In the case of audio deepfakes, spectrogram-based analysis combined with CNN models has proven effective in identifying synthetic speech patterns.

Despite these advancements, most existing systems operate independently within their respective domains. The integration of multiple modalities into a unified framework remains a challenge due to differences in data representation, computational complexity, and model compatibility. The proposed system addresses these challenges by combining multiple deep learning models into a single architecture, enabling comprehensive analysis and improved detection accuracy.

III. PROBLEM STATEMENT

The rapid evolution of deepfake generation techniques has created significant challenges in detecting manipulated content. Existing detection systems face several limitations that reduce their effectiveness in real-world scenarios. Firstly, most traditional systems rely on manual verification processes, which are time-consuming and impractical given the massive volume of online content. Secondly, many automated detection approaches focus on a single modality, such as text or images, and fail to consider the combined influence of multiple media types.

Additionally, modern deepfake techniques produce highly realistic outputs that closely resemble genuine content, making detection increasingly difficult. The lack of real-time processing capabilities further limits the usability of existing systems in applications such as social media monitoring and cybersecurity. Therefore, there is a critical need for a robust, scalable, and multimodal detection system capable of identifying deepfakes across various media formats with high accuracy and efficiency.

IV. PROPOSED SYSTEM

System Architecture

The proposed system adopts a multimodal deep learning architecture designed to process and analyze text, image, audio, and video inputs simultaneously. The architecture follows a layered approach consisting of six major layers: User Interface Layer, Input Processing Layer, Preprocessing Layer, Feature Extraction Layer, Multimodal Fusion Layer, and Output Layer.

A. Text Analysis Module

The text analysis component uses a pretrained BERT model with an attention mechanism to understand the semantic context of the input text. The model processes tokenized text and generates contextual embeddings used for classification. This approach enables the detection of subtle linguistic patterns associated with misinformation.

B. Image Analysis Module

For image-based deepfake detection, the system utilizes DenseNet121, a deep convolutional neural network known for its efficient feature propagation and reduced vanishing gradient problem. The model is trained to identify visual inconsistencies such as unnatural textures, facial distortions, and GAN-generated artifacts.

C. Audio Analysis Module

The audio processing module converts input audio signals into Mel spectrograms, which represent frequency variations over time. These spectrograms are analyzed using CNN models to detect anomalies in speech patterns, tone, and frequency distributions that indicate synthetic audio.

D. Video Analysis Module

Video data is processed by extracting frames and applying Eulerian Video Magnification (EVM) to amplify subtle facial movements. Feature extraction is performed using ResNext, followed by LSTM networks that capture temporal dependencies across frames. This allows the system to detect inconsistencies in facial expressions and motion patterns.

E. Multimodal Fusion and Classification

The outputs from all modules are combined using a fusion layer, which integrates features from different modalities. A final classification layer with a softmax function determines whether the content is real or fake and assigns a confidence score.

F. RESULTS AND DISCUSSION

The proposed system was implemented using Python as the primary programming language with the Flask framework for backend deployment. The system was evaluated using benchmark datasets including Celeb-DF and FaceForensics++ for video and image detection, ASVspoof for audio spoofing detection, and FakeNewsNet and LIAR for text-based misinformation analysis.

The BERT-based text module achieved high accuracy in detecting linguistically manipulated content by leveraging contextual embeddings and attention mechanisms. The DenseNet121 image module successfully identified facial distortions and GAN-generated artifacts across standard benchmark datasets. The audio CNN module demonstrated strong performance in distinguishing synthetic speech using Mel spectrogram analysis. The video analysis pipeline combining ResNext and LSTM networks effectively captured temporal inconsistencies in facial expressions across multiple frames.

Discussion

The results highlight the effectiveness of a multimodal, unified deep learning approach for deepfake

detection. By combining complementary information from text, image, audio, and video modalities through fusion layers, the system achieves improved robustness compared to single-modality approaches. The real-time web deployment ensures practical usability for applications in social media monitoring, digital forensics, and cybersecurity.

System Strengths

One of the primary strengths of the proposed framework lies in its multimodal architecture. The system processes diverse media types simultaneously, enabling comprehensive detection that single-modality systems cannot match. The use of pretrained models such as BERT and DenseNet121 reduces training overhead while maintaining high accuracy. GPU acceleration and efficient data handling ensure real-time performance suitable for practical deployment.

Limitations

Despite its successes, several limitations remain in the current prototype. The system requires significant computational resources for real-time multimodal processing. The detection accuracy may degrade against highly advanced next-generation deepfake generation techniques. Additionally, the system's performance depends on the quality and diversity of training datasets, and further work is needed to generalize across emerging deepfake types.

V. CONCLUSION

This paper presents a comprehensive and efficient approach for detecting deepfake content using a multimodal artificial intelligence framework. The proposed system integrates advanced models such as BERT for text analysis, convolutional neural networks for image and audio processing, and LSTM-based architectures for video analysis, enabling accurate detection across multiple data formats. The implementation of a unified architecture combining text, image, audio, and video analysis significantly improves the system's capability to detect complex and high-quality deepfakes.

By leveraging multimodal fusion techniques, the system enhances detection accuracy and robustness compared to traditional single-modality approaches. The deployment as a web-based application ensures accessibility and practical usability for real-time deepfake detection, suitable for applications in social media monitoring, digital forensics, cybersecurity, and content verification. With further advancements and integration of emerging technologies, the system has the potential to play a vital role in preserving the authenticity and integrity of digital information.

VI. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Sir Issac Newton College of Engineering and Technology, for providing the necessary guidance, resources, and technical support throughout the development of this project. The authors also appreciate the encouragement and support received from the faculty members and the institution, which made the successful completion of this research work possible.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), 2014.

[2] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.

[4] Y. Li, M. Chang, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE International Conference on Computer Vision (ICCV), 2019.

[6] X. Zhang, S. Karaman, and S. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," IEEE International Workshop on Information Forensics and Security (WIFS), 2019.

[7] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2020.

[8] ASVspoof Consortium, "ASVspoof Challenge: Automatic Speaker Verification Spoofing and Countermeasures," 2021.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," International Conference on Learning Representations (ICLR), 2015.