AI DEEP FORGE: AN INTEGRATED UNIMODEL FOR DIGITAL FORENSICS IN AI-GENERATED CONTENT

Project Reference No.: 49S_BE_3658

College : Bearys Institute of Technology, Mangaluru
Branch : Department of Artificial Intelligence and Data Science
Guides : Dr. Mehaboob Mujawar
Prof. Soudha N
Students : Ms. Sumaiya Naseer Husen Sarwad
Ms. Asiyamath Azeema
Ms. Ayishath Safa
Ms. Fathima Anna

Keywords:

Synthetic media, deepfake detection, multimodal forensic analysis, AI-generated content, text detection, image detection, audio detection, video detection.

Introduction:

The rapid advancement of Generative Artificial Intelligence (AI) has significantly transformed the way digital content is created and consumed. Modern AI models can now generate highly realistic text, images, audio, and videos that are often difficult to distinguish from genuine human-created content. While these technologies offer many beneficial applications, they also introduce serious risks when misused. Synthetic media can be exploited for spreading misinformation, deepfake creation, identity theft, financial fraud, defamation, and social manipulation.

The increasing presence of such AI-generated content on social media and digital platforms has created major concerns regarding digital trust, authenticity, and cybersecurity. Most users lack the technical expertise to reliably differentiate between authentic and manipulated content, making them vulnerable to deception. This has created an urgent need for intelligent forensic tools capable of verifying media authenticity quickly and accurately.

To address this challenge, the project AI Deep Forge proposes a unified multimodal digital forensic framework that can authenticate content across multiple media formats within a single system. The framework supports the analysis of text, image, audio, and video inputs, enabling users to upload suspicious content and receive authenticity predictions in real time.

The system combines machine learning and deep learning models, such as TF-IDF based classifiers for text and CNN-based architectures for image, audio, and video detection. A Flask-based web application is used to integrate all detection modules into a user-friendly

interface. This centralized approach improves usability, scalability, and practical deployment for real-world forensic applications.

By providing a single platform for multimodal fake media detection, AI Deep Forge aims to strengthen digital trust, cyber forensics, and secure online communication.

Objectives:

1. To design and develop a multi-modal AI-based system capable of detecting and classifying media as real or AI-generated.
2. To implement individual analyzers for text, image, audio, and video using appropriate machine learning and deep learning techniques.
3. To build a Flask-based web application that allows users to upload and analyze different types of media through an interactive interface.
4. To ensure the system provides accurate, reliable, and explainable results for each media type.
5. To create a training and retraining module that enables model improvement with new datasets over time.

Methodology:

1. Input Media / Dataset

The project follows a descriptive and experimental research methodology using curated datasets containing both real and AI-generated media samples. Separate datasets are collected for text, image, audio, and video modalities from reliable sources such as Kaggle repositories and benchmark deepfake datasets. These datasets include human-written and AI-generated text, real and synthetic images, human voice and AI-generated audio, as well as genuine and manipulated videos. The collected media is organized into AI (fake) and Real (genuine) categories to support supervised learning and evaluation.

2. Preprocessing

After data collection, preprocessing is performed to standardize all input formats before model training and inference. Text data undergoes tokenization, normalization, and cleaning to remove noise and unnecessary symbols. Images are resized into fixed dimensions and normalized for uniform pixel distribution. Audio files are converted into a standard sampling rate and processed using MFCC extraction techniques. Video files are handled through frame extraction, where important frames are selected and resized for further analysis. This stage improves consistency and helps reduce unwanted variations in the data.

3. Feature Extraction

Once preprocessing is completed, relevant features are extracted from each media type. Text inputs are transformed into numerical representations using TF-IDF and NLP-based embeddings. Image data uses CNN layers to automatically capture edges, textures, and artifact-based inconsistencies. Audio inputs are converted into spectrograms, MFCCs,

spectral centroid, and zero-crossing rate features to identify synthetic voice patterns. For videos, temporal and frame-level features are extracted to capture manipulation traces and visual inconsistencies across frames.

4. Detection Models

The extracted features are then passed into specialized machine learning and deep learning detection models. Text classification is performed using TF-IDF based classifiers, while image, audio, and video modules rely on CNN-based neural networks for accurate prediction. Each model is trained separately using its corresponding dataset and optimized to distinguish between real and AI-generated content. The complete backend is implemented using the Flask framework, integrating all modality-specific analyzers into a single unified system.

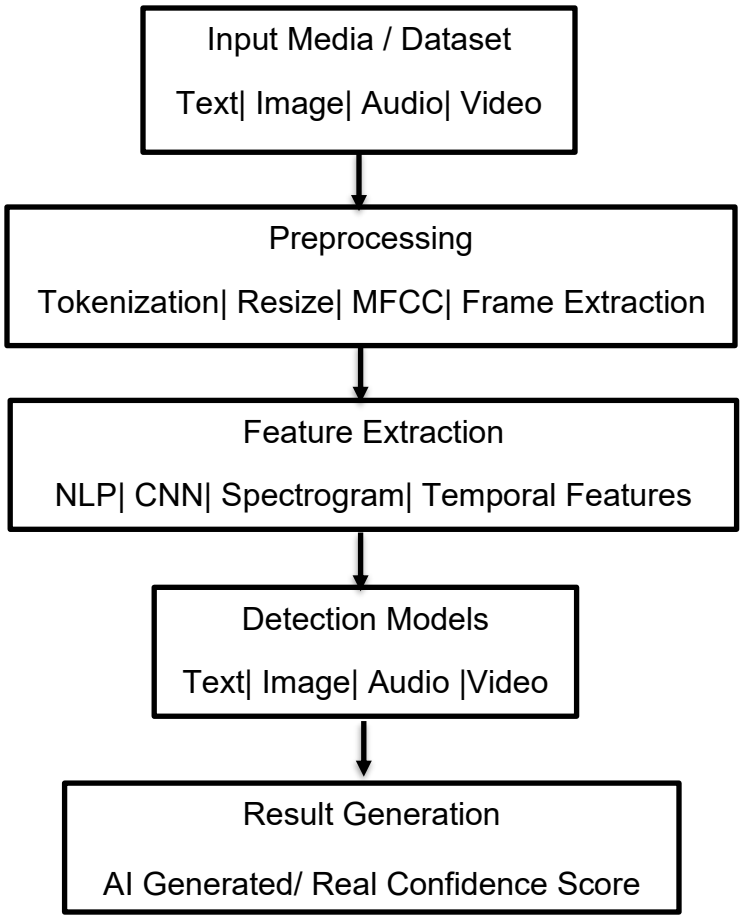

Figure 1: Methodology of project

5. Result Generation

After classification, the system generates a final forensic verdict based on the selected media analyzer. The prediction outputs are converted into a human-readable result, indicating whether the uploaded content is AI-generated or real, along with a confidence score showing the certainty of the model. Additional forensic indicators such as suspicious

artifacts, spectral anomalies, or frame inconsistencies are highlighted wherever applicable. The final result is displayed on the respective Flask web interface page, providing users with an accurate and interpretable authenticity report.

Result and Conclusion:

In conclusion, AI Deep Forge presents a unified and practical multimodal forensic platform for detecting whether digital content is real or AI-generated. The system successfully integrates text, image, audio, and video analysis within a single web-based framework, making the process of authenticity verification simple and accessible. By combining multiple detection pipelines into one centralized system, the project addresses the major limitation of fragmented single-modality forensic tools currently available.

The framework is capable of generating clear classification outputs, confidence scores, and modality-specific forensic indicators, which help users understand the basis of the authenticity verdict. Through its Flask-based interface, the platform offers a user-friendly experience where even non-technical users can upload suspicious media and obtain quick forensic insights. This improves the practical usability of AI-based digital forensic solutions in real-world scenarios.

Another key strength of the project lies in its modular architecture, where each analyzer for text, image, audio, and video functions independently while remaining part of the same unified system. This design ensures scalability, maintainability, and future extensibility, allowing improvements or model upgrades without affecting the complete framework.

The project also demonstrates how machine learning and deep learning techniques can be effectively applied to strengthen digital trust, cyber forensics, and online safety. As synthetic media continues to evolve, AI Deep Forge provides a scalable, efficient, and future-ready solution for combating misinformation, fraud, and manipulation.

In conclusion, the project offers a reliable approach for digital authenticity verification and serves as a strong foundation for advanced multimodal forensic research and deployment.

References:

1. [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680, 2014.
2. [2] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Security (WIFS), pp. 1–7, 2018.
3. [3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1–11, 2019.
4. [4] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4401–4410, 2019.

5. [5] X. Zhang et al., “Detecting AI-generated fake images in the wild,” arXiv preprint arXiv:2008.12399, 2020.
6. [6] C. Yu, P. Chen, J. Tian, J. Liu, J. Dai, X. Wang, Y. Chai, S. Jia, S. Lyu, and J. Han, “A unified framework for modality-agnostic deepfakes detection,” arXiv preprint arXiv:2307.14491, 2023. :contentReference[oaicite:0]{index=0}
7. [7] K. Gandhi, P. Kulkarni, T. Shah, P. Chaudhari, M. Narvekar, and K. Ghag, “A multimodal framework for deepfake detection,” arXiv preprint arXiv:2410.03487, 2024. :contentReference[oaicite:1]{index=1}
8. [8] A. H. Soudy, O. Sayed, H. Tag-Elser, et al., “Deepfake detection using convolutional vision transformers and convolutional neural networks,” Neural Computing and Applications, vol. 36, pp. 19759–19775, 2024.
9. [9] X. Zhang, J. Chu, J. Zhao, Y. Jiang, X. Yang, L. Jin, C. Zhang, and X. Li, “ERF-BA-TFD+: A multimodal model for audio-visual deepfake detection,” arXiv preprint arXiv:2508.17282, 2025. :contentReference[oaicite:2]{index=2}
10. [10] F. Chollet, Deep Learning with Python. Shelter Island, NY, USA: Manning Publications, 2017.

Future Scope:

The future scope of this project includes:

1. Integration of advanced transformer-based and cross-modal fusion models to improve detection accuracy across text, image, audio, and video.
2. Expansion to multiple languages, additional media types, and cloud-based deployment for large-scale real-time forensic analysis.
3. Incorporation of Explainable AI (XAI) for transparent forensic outputs and extension to applications such as newsroom verification, social media moderation, cybersecurity, academic integrity checks, and law-enforcement support.
4. Development of a mobile application interface to enable instant authenticity checks directly from smartphones and portable devices.
5. Integration of blockchain-based evidence tracking and secure audit logs to preserve the integrity and traceability of forensic reports.
6. Enhancement of the system with real-time social media monitoring APIs for automatic detection of synthetic media spreading across online platforms.