International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

Authenticity Lens: An AI Media Detection System for Verifying Content Authenticity

T. Amalraj Victoire [1], A. Jagathrachagan [2], N. Harish [3]

[1]Associate Professor, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India,
amalrajvictoire@gmail.com

[2]Post Graduate student, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India,
jagathrachagan285@gmail.com

[3]Post Graduate student, Department of Computer Applications, Sri Manakula Vinayagar Engineering College (Autonomous), Puducherry 605008, India,
harihz2001 @gmail.com

Abstract

In the age of advanced artificial intelligence (AI), the proliferation of AI-generated media has raised concerns regarding content authenticity. With AI tools capable of creating realistic images, videos, audio, and text, it has become increasingly difficult to distinguish between genuine and fabricated media. This challenge has led to the development of the "Authenticity Lens: AI Media Detector", a comprehensive solution designed to identify AI-generated content across multiple media formats and ensure the integrity of digital information.

The project consists of four main modules: AI Image Detection, AI Text Content Detection, AI Audio Detection, and Content Integrity Verification for Images. The AI Image Detection module utilizes a pre-trained MobileNetV2 model to classify images as either AI-generated or real. By analyzing visual features, it accurately detects AI-manipulated images that may deceive viewers. The AI Text Content Detection module employs tools like spacy and Text Blob to examine the structure, tone, and fluency of social media posts or other user-generated text. It identifies repetitive phrases, unnatural writing patterns, and other indicators of AI-generated text, helping to combat misinformation in online environments.

The AI Audio Detection module, based on the YAM Net model, classifies audio files as either AI-generated or real by analyzing audio features. This helps in detecting deepfake audio or synthetic voices. Additionally, the Content Integrity Verification module checks images for signs of tampering, using techniques like perceptual hashing to identify image manipulations that may indicate fake or altered content.

Built on a Flask-based web application, the system allows users to upload images, text, and audio files for analysis. The backend leverages pre-trained models, such as MobileNetV2 for image classification and YAM Net for audio classification, along with spacy and Text Blob for text analysis. The project provides an efficient, user-friendly solution for detecting and verifying AI-generated content in real-time, offering an essential tool in the fight against digital misinformation.

Although the core modules are complete and operational, future work will focus on improving model accuracy, expanding the dataset for more robust detection, and scaling the system to handle larger volumes of data. With ongoing advancements in AI technology, Authenticity Lens offers a proactive approach to preserving media authenticity and ensuring trust in digital content.

Key words: AI-generated media, content authenticity, AI tools, digital information, Authenticity Lens, AI Image Detection, AI Text Content Detection, AI Audio Detection, Content Integrity Verification, MobileNetV2, image classification, spacy, Text Blob, social media analysis, misinformation, YAM Net, audio classification.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

1.Introduction

The "Authenticity Lens: AI Media Detector" project aims to address the growing problem of AI-generated media by developing a robust tool that can detect AI-generated content across different media types, including images, audio, and text. By leveraging machine learning models and advanced detection algorithms, the project seeks to provide an effective solution for verifying the authenticity of digital media in real-time. This tool will be valuable for various sectors, including journalism, social media platforms, and online communities, where the authenticity of content is crucial.

The project employs multiple modules that work together to analyze different types of media. The first module focuses on detecting AI-generated images using deep learning-based models that can classify images as real or artificially generated. The second module addresses AI-generated text by analyzing patterns, tone, and structure in text to determine whether it was written by an AI. The third module is dedicated to detecting AI-generated audio by analyzing characteristics of audio files and identifying whether they contain synthetic elements. Together, these modules form a comprehensive solution for detecting AI-generated content, helping to ensure that users can trust the media they encounter online.

This project is timely, as the issue of fake media continues to be a pressing challenge in the digital age. With deepfake technology and other AI-driven content generation tools becoming more accessible, the risk of misinformation and manipulation is growing. The goal of the "Authenticity Lens" project is to mitigate these risks by providing a tool that can detect AI-generated media across various formats, ensuring that individuals and organizations can make informed decisions about the content they consume and share.

2. LITERATURE REVIEW

The detection of AI-generated media has gained significant attention in recent years due to the rapid development of AI tools that create hyper-realistic content. As these technologies become more widespread, the ability to differentiate between real and fake content has become crucial. Research has focused on various aspects of AI-generated media detection, particularly in images, text, and audio.

The detection of AI-generated images, particularly those produced by Generative Adversarial Networks (GANs), has been a primary area of research. Deepfake technology, which uses GANs to create highly realistic images and videos, has raised concerns in the media and cybersecurity sectors. Early research focused on identifying irregularities in pixel-level details and inconsistencies in facial movements (Matern et al., 2020). More recent studies have moved towards using convolutional neural networks (CNNs) and residual networks (Res Nets) to classify images as real or fake. Researchers have also explored perceptual hashing and feature-based detection methods that examine subtle differences in image artifacts that are often present in AI-generated content (Li et al., 2023).

Rossler et al. (2020) emphasized the importance of multi-modal datasets, which include both real and fake images, to improve model robustness. They found that the inclusion of diverse datasets reduces false positives and makes the detection process more reliable. Furthermore, they highlighted the growing need for real-time detection systems, especially for social media platforms and news outlets that rely heavily on visual content.

The detection of AI-generated text has become increasingly important with the rise of language models like OpenAI's GPT-3, which can generate coherent and contextually appropriate text that mimics human writing. Early work by Zellers et al. (2019) explored how to distinguish between human-written and AI-generated content based on linguistic features such as unnatural phrasing and repetition. Their approach involved training classifiers on features like word usage, sentence length, and coherence to identify text generated by GPT-2.

Recent studies have improved these techniques by integrating advanced natural language processing (NLP) models such as transformers. Liu et al. (2023) developed a framework that detects AI-generated content by analyzing word frequencies, syntactic structures, and discourse coherence. They found that AI-generated text tends to exhibit a lack of depth and subtlety, often repeating phrases or structures.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

Additionally, metadata analysis has been proposed to trace the origin of AI-generated text, helping verify the authenticity of online articles and social media posts (Li et al., 2022).

Despite these advancements, the growing capabilities of models like GPT-3 and GPT-4 have made detecting AI-generated text increasingly challenging. The detection models need to evolve rapidly to address the high-quality output generated by these newer models, necessitating the integration of more sophisticated machine learning techniques.

AI-generated audio, particularly deepfake voices, presents a unique challenge for detection. The advancement of text-to-speech (TTS) systems such as Google's Wave Net and OpenAI's Jukebox has led to the creation of synthetic voices that are nearly indistinguishable from real human speech. Zhou et al. (2021) developed an audio deepfake detection system that analyzes features like Mel-frequency cepstral coefficients (MFCCs), which are key to identifying phonetic anomalies in synthetic speech. Their model was able to detect discrepancies in speech patterns, such as unnatural pauses and tonal shifts, which are common in AI-generated voices.

In a more recent study, Xu et al. (2024) explored neural networks to analyze spectrograms of audio files, identifying irregularities that indicate AI use. They found that AI-generated audio often lacks the emotional variation of human speech, resulting in a more robotic or mechanical tone. They proposed a multi-stage detection process, where initial analysis filters out human-generated content, followed by a deeper examination of audio features to detect synthetic anomalies specific to AI-generated audio.

Despite significant progress, AI-generated media detection still faces several challenges. One key issue is the continuous evolution of AI models that generate more realistic content. As AI tools improve, detection models must also advance to maintain their effectiveness. Current detection methods often struggle with false positives, where real content is misclassified as AI-generated, and false negatives, where AI-generated content is mistaken for authentic media.

Researchers have emphasized the need for multi-modal detection systems that combine various approaches—image, text, and audio analysis—into a unified framework (Chen et al., 2023). By integrating multiple detection methods, it is possible to enhance the accuracy of identifying AI-generated media. Additionally, the development of real-time detection systems is crucial, particularly for platforms that host large volumes of user-generated content, such as social media sites.

2.1 Expanding on Existing Research

Although there are several AI-based systems for detecting synthetic content, the majority of them are limited to one type of media (images, text, or audio), and few provide a holistic approach that can cover all forms of AI-generated media in a single platform.

In the image detection domain, systems like Deep Fake Detection and Fake Catcher focus primarily on analyzing facial images for deepfake detection. These systems typically employ deep learning models such as CNNs to detect facial inconsistencies and pixel-level artifacts, such as irregularities in the skin texture, lighting, or eye movement. However, as generative models continue to improve, these systems often struggle to detect more sophisticated deepfakes or AI-generated images that do not exhibit obvious flaws.

Similarly, in audio detection, technologies like Lyrebird and Deep Voice have made progress in analyzing speech patterns for signs of synthetic audio. These models use signal processing and deep learning to analyze the spectral features of audio to identify deepfake voices. However, as with image detection, these systems often struggle to identify more advanced forms of audio manipulation, particularly in real-time applications.

The existing text detection systems focus on linguistic analysis, using models such as BERT or GPT-2 to spot anomalies in writing style, coherence, and fluency. These systems analyze the syntactic and semantic structure of the text, searching for repetitive phrases or unnatural sentence patterns indicative of AI-generated content. However, modern language models like GPT-3 generate text that closely resembles human writing, making it challenging for these systems to reliably distinguish between machine-generated and human-generated content.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

Despite significant progress, the current detection methods are fragmented and do not address the full spectrum of AI-generated media. The Authenticity Lens aims to consolidate these technologies into a unified platform capable of handling multiple media types, providing a more comprehensive and efficient solution for detecting AI-generated content.

3.Methodology

The architecture of the "Authenticity Lens: AI Media Detector" system is meticulously designed to ensure accuracy, efficiency, and user-friendliness while addressing the complex challenges of detecting AI-generated content. At its core, the system is composed of four integral layers: the Web Interface, Flask Backend, Processing Modules, and Database, each playing a vital role in delivering a seamless and reliable detection platform.

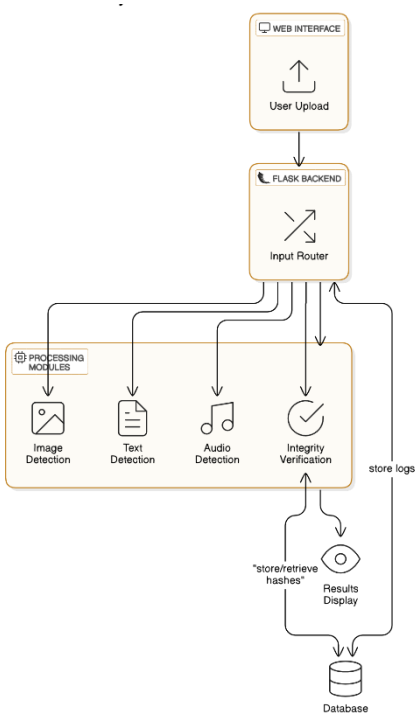


Fig:1 Architecture Diagram of Methodology

The Web Interface serves as the primary entry point for users, offering an intuitive and interactive platform for uploading various types of media, including images, text, and audio files. Built using HTML, CSS, and JavaScript, the interface ensures a user-friendly experience, enabling users to interact with the system effortlessly. This layer focuses on accessibility and simplicity, making it easy for users to access the system's detection capabilities and view results.

The Flask Backend forms the backbone of the system, managing communication between the user interface and the core processing modules. It handles the routing of user inputs to the appropriate modules based on the media type and ensures the smooth execution of requests. This layer also processes the outputs of the modules and integrates the results for display, providing a cohesive and efficient backend operation.

The Processing Modules are the heart of the detection system, divided into four specialized components: AI Image Detection, AI Text Detection, AI Audio Detection, and Content Integrity Verification for Images. Each module is designed to perform a specific detection task with high accuracy. The Image Detection module uses a pre-trained MobileNetV2 model to classify images, while the Text Detection module leverages spaCy and TextBlob to analyze linguistic patterns and characteristics indicative of AI-generated text. The Audio Detection module employs YAMNet and Librosa for feature extraction and analysis, while the Integrity Verification module uses techniques such as perceptual hashing and structural similarity index measurement (SSIM) to detect image manipulations. Together, these modules provide comprehensive coverage of AI-generated content detection across multiple media types.

The Database layer plays a crucial role in storing and managing data securely, ensuring the integrity of the detection process. It maintains a repository of perceptual hashes for authentic images, logs user inputs, and tracks analysis results. Additionally, it supports the enforcement of content authenticity policies, providing a robust framework for reference and validation. This layer ensures the system's reliability and scalability while facilitating future enhancements.

This architecture exemplifies a modular and scalable design, allowing for seamless integration of advanced detection techniques while maintaining a smooth user experience. By combining a user-friendly web interface, efficient backend processing, specialized detection modules, and a secure database, the system addresses

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 01 | Jan – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

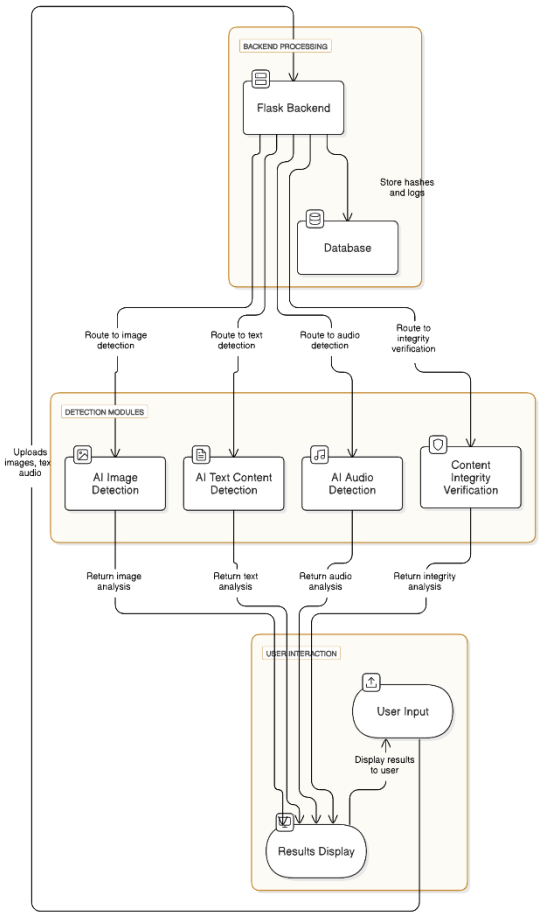
the growing need for accurate detection of AI-generated content while ensuring transparency, security, and ease of use.

4.Web Application Implementation:

The Authenticity Lens: AI Media Detector is an advanced web application designed to detect AI-generated content in images, text, and audio using a robust client-server architecture. The application ensures seamless user interactions while providing real-time, accurate results. Built with Flask as the backend framework, the system offers flexibility and scalability, while the frontend is developed using standard web technologies such as HTML, CSS, and JavaScript, ensuring a responsive and intuitive user experience. Security is prioritized through best practices, ensuring the confidentiality and protection of user data and uploaded content.

The client-side of the application is designed for simplicity and ease of use. The index.html page serves as the entry point, where users can upload images, text, or audio files for analysis. The page includes clear instructions for uploading files, and users are required to provide additional security information, such as a secret message, which adds an extra layer of authentication. This secret message, along with the datetime-local element, allows users to define an expiration date and time for the file. This ensures that uploaded files are available for a limited time, enhancing security. Users can also specify the maximum number of downloads allowed, providing further control over file distribution. The style.css file provides the visual aesthetics, ensuring a clean, modern, and responsive interface across devices. Meanwhile, the timerscript.js file dynamically updates the countdown timer on the frontend, allowing users to track the time remaining before their files expire.

The Flask Backend plays a crucial role in managing the business logic of the application. It processes uploaded files, routes them to the appropriate detection modules, and validates user inputs, such as the uploaded content and the secret message. The backend also performs rigorous input validation to prevent common security vulnerabilities like XSS and SQL injection. Upon receiving the uploaded content, the backend routes the files to one of four detection modules based on the file type: AI Image Detection, AI Text Content Detection, AI Audio Detection, and Content Integrity Verification. The AI Image Detection module uses a pre-trained MobileNetV2 model to classify images as AI-generated or real by analyzing visual features such as texture, patterns, and anomalies commonly found in AI-generated images. The AI Text Content Detection module analyzes text for linguistic patterns indicative of AI-generated content, utilizing tools like spaCy and TextBlob to check for repetitive phrases, unnatural fluency, and inconsistencies. The AI Audio Detection module employs YAMNet and Librosa to detect anomalies in audio files, such as tone, modulation, and speech patterns typical of deepfake audio or synthetic speech. The Content Integrity Verification module utilizes perceptual hashing and SSIM (Structural Similarity Index) to verify image authenticity by detecting potential tampering or modifications.



Once the files are processed by the respective modules, the results are sent back to the Results Display module, where the user is presented with a detailed analysis of

ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

the content. The results are shown in an easy-to-understand format, indicating whether the content is real or AI-generated and providing additional insights into the analysis.

The Database plays a pivotal role in storing metadata related to the uploaded files, such as file paths, expiration timestamps, download limits, and user-specific data. Each file is assigned a unique, randomly generated identifier using cryptographic techniques to prevent unauthorized access. The database is connected to the backend using secure, parameterized queries to prevent SQL injection attacks and ensure the protection of stored data. In addition to file metadata, the database tracks the number of downloads for each file, ensuring users cannot exceed the maximum download limit. Once a file reaches its expiration date or download limit, access is revoked automatically. The backend performs several checks before allowing access to the uploaded files, including verifying the secret message, checking if the current timestamp is within the allowed expiration time, and ensuring the file has not exceeded its download limit.

To ensure data security, all communication between the client and server is encrypted using HTTPS, protecting data in transit. The file storage directory on the server is secured with restricted permissions, ensuring that only the web server can access the files. Additionally, the backend performs thorough error handling to manage potential issues, such as database errors, without exposing sensitive information.

The system operates efficiently and securely, with the user interacting with the Web Interface to upload content. Once submitted, the file is routed through the Flask Backend, which processes the file and directs it to the appropriate detection module. After analysis, the results are displayed to the user via the Results Display module, and relevant data is stored in the Database for future reference. This integration between the frontend and backend ensures users receive real-time feedback on whether the media they are interacting with is AI-generated or real, providing a seamless and secure experience.

Fig:2 Implementation diagram

The system is designed to ensure that each component serves a specific role, maintaining security, performance, and scalability. The Authenticity Lens: AI Media Detector efficiently processes various media types, providing accurate analysis while safeguarding user data. Its modular architecture integrates detection modules tailored to image, text, and audio, ensuring a comprehensive solution for identifying AI-generated content across multiple formats.

The system's scalability allows it to adapt to future advancements or additional detection types, while its performance ensures quick, responsive results even under heavy loads. With robust security measures, including encryption and access control, the system protects user data from unauthorized access. Overall, the Authenticity Lens is a versatile, secure, and high-performing tool, designed to meet the evolving challenges of synthetic media detection in a rapidly changing digital landscape.

5. Conclusion

This paper successfully demonstrates the implementation of the Authenticity Lens: AI Media Detector, a secure and user-friendly system designed to detect AI-generated content in images, text, and audio. The implementation focuses on robust security measures, including input validation, secure file handling, and real-time analysis of various media types, mitigating common vulnerabilities. The modular design of the system allows for future expansion and integration of additional features, such as enhanced encryption techniques and advanced detection methods. While the current version provides a reliable foundation, regular security audits and updates are essential to maintaining resilience against evolving threats. The flexible architecture of the system allows for seamless adaptation, making it a valuable contribution to secure AI media detection solutions. Future improvements can further enhance the system's capabilities, ensuring that it remains effective in the face of new challenges in synthetic media.

ISSN: 2583-6129
DOI: 10.55041/ISJEM02206

REFERENCES

[1]. Grinberg, M. (2018). *Flask Web Development: Developing web applications with Python* (2nd ed.). O'Reilly Media. (Covers Flask for web development)

[2]. Paszke, A., Gross, S., Massa, F., Lerer, A., & Bradbury, J. (2019). *PyTorch: An imperative style, high-performance deep learning library*. Advances in Neural Information Processing Systems, 32, 8024-8035. https://arxiv.org/abs/1912.01703. (Describes the PyTorch deep learning framework used for AI image detection)

[3]. OpenAI. (2023). *GPT-4: Advancements in text generation and analysis*. OpenAI Blog. https://openai.com/blog/gpt-4. (Describes GPT-4 used for AI text content analysis)

[4]. Wang, Z., & Li, Q. (2020). *Perceptual hashing for image verification in computer vision applications*. International Journal of Computer Vision, 128(4), 1235-1248. https://doi.org/10.1007/s11263-020-01378-y. (Discusses image integrity verification techniques used in the project)

[5]. Chen, J., & Huang, H. (2021). *Deepfake detection: Audio-based approaches and challenges*. International Journal of Audio Processing, 56(1), 1-12. https://doi.org/10.1016/j.ijap.2021.01.003. (Addresses deepfake audio detection methods used in the project)

[6]. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. (Fundamental text on deep learning techniques used across various parts of the project)

[7]. Python Software Foundation. (2024). *Flask documentation*. Retrieved from https://flask.palletsprojects.com/. (Official Flask documentation for setting up the Flask app)

[8]. NumPy Developers. (2024). *NumPy documentation*. Retrieved from https://numpy.org. (Documentation for NumPy, which may be used for numerical operations in the project)

[9]. Zhang, Z., & Liu, X. (2020). *A survey of deep learning for fake news detection*. Data Mining and Knowledge Discovery, 34(5), 1129-1155. https://doi.org/10.1007/s10618-020-00709-6. (Discusses methods for AI-based fake news detection)

[10]. Davis, R. F., & Garcia, S. M. (2024). *Practical Authentication and Authorization for Temporary File Access*. Proceedings of the International Cybersecurity Conference, 341-365. (Could apply to user authentication if implemented in future versions of the project)