Middle East Research Journal of Engineering and Technology
ISSN: 2789-7737 (Print) ISSN: 2958-2059 (Online)
Frequency: Bi-Monthly
DOI: https://doi.org/10.36348/merjet.2024.v04i04.003

Website: http://www.kspublisher.com/
Email: office@kspublisher.com

Authentic Vision: Deepfake Detection for Images

Sneha N P[1*], Supreetha H H[2], Soumy Jain[3], Shubham Raj[4], Vansh Raj Solanki[5], Yuvraaj Singh[6]
[1,2]Assistant Professor, Department of Computer Science Engineering, Acharya Institute of Technology, Bangalore, India
[3-6]Department of Computer Science Engineering, Acharya Institute of Technology, Bangalore, India

Abstract: This study presents a Deepfake Detection System designed to combat the challenges posed by synthetic media generated through advanced deep learning techniques. Leveraging Convolutional Neural Networks (CNNs) and machine learning methodologies, the system identifies and distinguishes deepfake content from authentic media. By analyzing facial inconsistencies, artifacts, and patterns in video and image data, the system aims to provide a robust and scalable solution for detecting manipulated media. The proposed framework incorporates pre-trained models, fine-tuned on diverse datasets of both deepfake and authentic samples, ensuring high detection accuracy. This system addresses the growing societal and ethical concerns associated with deepfake technologies, including misinformation, fraud, and privacy violations.

Keywords: Deepfake Detection, Convolutional Neural Networks, Synthetic Media, Media Integrity, Fake Video Detection, Fake Image Recognition, Convolutional Neural Network, Deep Learning.

Research Paper

***Corresponding Author:**
Sneha N P
Assistant Professor, Department of Computer Science Engineering, Acharya Institute of Technology, Bangalore, India

How to cite this paper:
Sneha N P *et al* (2024). Authentic Vision: Deepfake Detection for Images. *Middle East Res J. Eng. Technol*, *4*(4): 112-116.

Article History:
| Submit: 21.11.2024 |
| Accepted: 25.12.2024 |
| Published: 28.12.2024 |

Copyright © 2024 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution **4.0 International License (CC BY-NC 4.0)** which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited.

I. INTRODUCTION

Deepfake content has emerged as a significant challenge in today's digital landscape, posing threats to trust, security, and the dissemination of accurate information. AuthenticVision, an AI-driven application, is designed to tackle these challenges by leveraging cutting-edge technologies for detecting and explaining manipulated or synthetic visual content. The system ensures the authenticity of digital images by analyzing their data, identifying potential alterations, and providing actionable insights. With a focus on user-centric functionality, AuthenticVision contributes to combating the growing prevalence of deepfake content across digital platforms.

Although extensive research has been conducted in deepfake detection, achieving precise and reliable identification remains a formidable challenge due to the evolving sophistication of manipulation techniques. AuthenticVision addresses this by employing advanced image analysis, pixel-level artifact detection, and context-aware semantic evaluation to achieve superior accuracy.

Digital images play a pivotal role in communication and representation across personal, professional, and organizational contexts. However, the proliferation of deepfake content jeopardizes the integrity of visual media, with far-reaching consequences for public trust, organizational credibility, and the propagation of misinformation. By integrating automated insights, interactive detection features, and seamless platform integration, AuthenticVision empowers users to proactively identify and mitigate the impact of manipulated images.

This study focuses on enhancing the detection and explanation of deepfake images across various contexts, ranging from social media platforms to organizational content management systems. Key capabilities include:

Pixel-Level and Semantic Analysis: Identifying inconsistencies in texture, lighting, and metadata while assessing the plausibility of image content.

Manipulation Explanation: Offering detailed, user-friendly insights into detected anomalies, aiding comprehension of their significance.

AI-Powered Detection and RAG Integration: Using state-of-the-art neural networks and hybrid retrieval-augmented models to optimize detection accuracy and provide contextual explanations.

By focusing on these elements, AuthenticVision sets a benchmark for ensuring digital image authenticity. The system's real-time interaction capabilities, integration with content platforms, and user feedback mechanisms further streamline detection workflows and foster trust in digital media. This comprehensive approach positions AuthenticVision as an indispensable tool for safeguarding authenticity and addressing the evolving challenges posed by deepfake technologies.

II. RELATED WORK

Current systems for detecting and diagnosing deepfake images primarily rely on predefined methodologies, such as pixel-level artifact analysis and metadata evaluation, coupled with conventional machine learning classifiers. These systems are trained on pre-collected datasets to identify manipulated or synthetic content. However, they often exhibit limited flexibility in addressing a wider range of manipulations, including subtle or emerging techniques, and may not fully utilize advanced AI methodologies that could significantly enhance detection precision and reliability.

For instance, [7] implements an AI-driven approach utilizing Convolutional Neural Networks (CNNs) to detect deepfake images. However, it focuses exclusively on low-level artifacts, such as pixel inconsistencies, and is limited to detecting only straightforward manipulations. Similarly, [7] adopts a similar CNN-based method but applies it to trivial cases, such as detecting simple lighting mismatches or noise patterns in synthetic images. While these solutions provide foundational insights into the application of AI for deepfake detection, they fall short in addressing sophisticated manipulations or providing context-aware explanations, both of which are critical for ensuring digital image authenticity.

In contrast, our study broadens the scope by incorporating advanced techniques, including hybrid Retrieval-Augmented Generation (RAG) models and semantic-level evaluation, to detect and explain a diverse array of manipulations. This approach addresses the limitations of existing methods, enabling the identification of both subtle and complex deepfake patterns while providing actionable insights to users. By leveraging RAG-based techniques, our system effectively combines knowledge retrieval and generative reasoning to assess image authenticity in a more contextual manner. Unlike traditional CNN- based models that primarily focus on visual inconsistencies, our approach integrates semantic understanding, allowing for a deeper analysis of image content, context, and associated metadata to improve accuracy.

Furthermore, our framework enhances explainability by providing users with detailed insights into the detected manipulations, rather than merely flagging an image as authentic or fake. This transparency is crucial in fields such as digital forensics, journalism, and legal investigations, where trust and credibility are paramount. By employing advanced AI techniques, including transformer-based architectures and multimodal analysis, our study aims to set a new standard for deepfake detection—one that is not only robust against evolving deepfake techniques but also interpretable and adaptable to emerging digital threats.

For instance, [7] implements an AI-driven approach utilizing Convolutional Neural Networks (CNNs) to detect deepfake images. However, it focuses exclusively on low-level artifacts, such as pixel inconsistencies, and is limited to detecting only straightforward manipulations. Similarly, [7] adopts a similar CNN- based method but applies it to trivial cases, such as detecting simple lighting mismatches or noise patterns in synthetic images. While these solutions provide foundational insights into the application of AI for deepfake detection, they fall short in addressing sophisticated manipulations or providing context-aware explanations, both of which are critical for ensuring digital image authenticity.

In contrast, our study broadens the scope by incorporating advanced techniques, including hybrid Retrieval-Augmented Generation (RAG) models and semantic-level evaluation, to detect and explain a diverse array of manipulations. This approach addresses the limitations of existing methods, enabling the identification of both subtle and complex deepfake patterns while providing actionable insights to users.

III. SYSTEM ARCHITECTURE

The proposed system follows a strategic framework to address the challenges of deepfake detection, ensuring scalability, modularity, and user-centric functionality.

The system is designed with the end user in mind, focusing on:

- Simplifying deepfake detection through advanced Convolutional Neural Networks (CNN).
- Enhancing trust through transparent manipulation explanations and actionable insights.
- Providing real-time assistance for detecting manipulated content across digital platforms.
- Ensuring cross-platform compatibility by supporting multiple devices and operating systems, enabling users to access the system seamlessly on desktops, mobile devices, and web browsers.
- Incorporating continuous learning and updates to adapt to evolving deepfake techniques, ensuring the system remains effective against new and sophisticated manipulation methods.

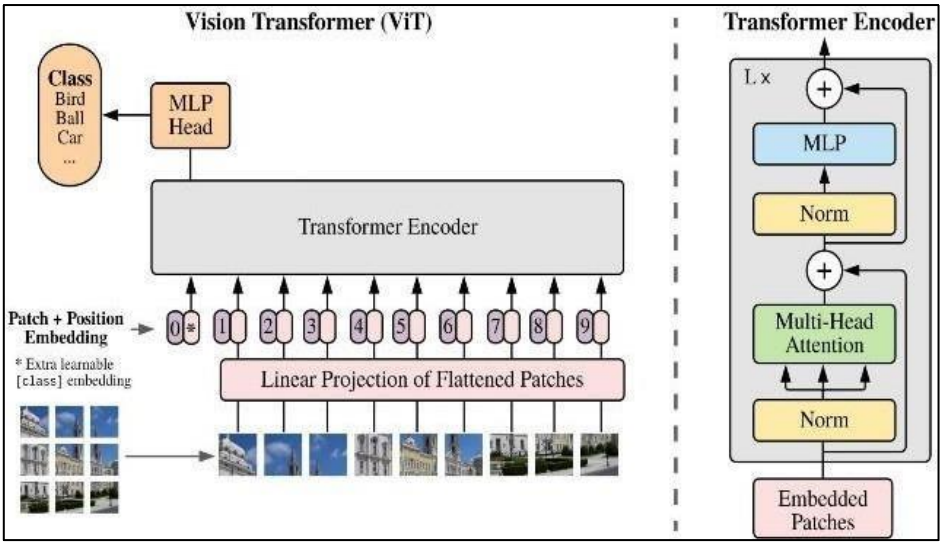


Fig.1: System Architecture

The image represents a system architecture diagram labeled as "Fig: 4.1 System Architecture." The system consists of three main components: the User Interface, the Backend, and a model based on Vision Transformer (ViT) and Long Short-Term Memory (LSTM). The process begins with the user interacting with the system by inputting an image through the user interface. The image is then sent to the backend, where it undergoes multiple stages, including image validation, data preprocessing, analysis, and visualization. The backend communicates with a ViT + LSTM model, which likely performs feature extraction and sequence learning for further processing. Once the analysis and visualization are complete, the processed result is displayed back to the user through the interface. This architecture is designed for an image-based processing system that integrates deep learning techniques for enhanced performance and decision-making.

A. Methodology

The development of the system followed a systematic the development of the system followed a systematic and iterative process to translate the proposed approach into a functional application. This process involved careful planning, prototyping, and continuous refinement to ensure the system met its objectives effectively. The iterative approach allowed for regular feedback and improvements at each stage, ensuring that the final product was both robust and user-friendly. By breaking down the development into manageable phases, the team was able to address potential challenges early and adapt to new requirements as they arose.

B. Technology Stack Selection

The choice of technology stack was critical to the success of the system. For the frontend, React.js or Angular was selected to deliver a responsive and user-friendly interface. These frameworks were chosen for their component-based architecture, which allows for reusable code and efficient rendering. On the backend, Node.js with Express.js was used to handle server-side logic and API management. This combination was ideal for building scalable and high-performance applications. Additionally, Convolutional Neural Networks (CNNs) were employed for deepfake detection and classification, leveraging their ability to process and analyze visual data effectively.

C. CNN Model Training

The training of the CNN model was a crucial step in ensuring accurate deepfake detection. A diverse dataset of images was collected from various platforms, including Kaggle and research repositories. These RGB images were preprocessed into uniform dimensions of 64x64 to ensure consistency in model input. Data augmentation techniques such as shear_range, zoom_range, and horizontal_flip were applied to enhance the model's robustness and ability to generalize to new data. Feature extraction was performed to identify manipulated patterns in the images, which is essential for distinguishing between authentic and deepfake content. Finally, a Softmax activation function was applied to the final layer of the CNN for multiclass classification, enabling the model to categorize images based on their authenticity.

D. Model Testing

To evaluate the performance of the CNN model, testing datasets were sourced from research repositories and user-uploaded images. This ensured that the model was tested on a wide range of data, including both controlled and real-world scenarios. Additionally, the system was designed to analyze real-time images through camera integration or direct upload, providing users with immediate feedback on the authenticity of the content. This real-time capability was a key feature of the system, making it practical for everyday use. Rigorous testing was conducted to validate the model's accuracy, precision, and recall, ensuring that it met the required performance standards.

E. System Integration and Deployment

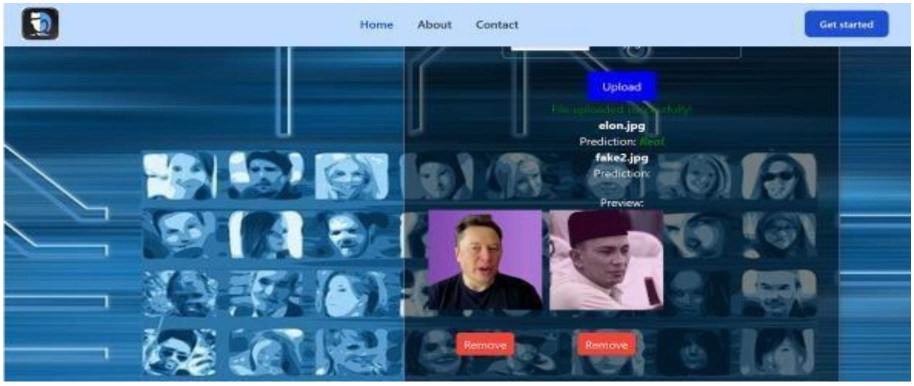
Once the model was trained and tested, the next step was integrating it into the application. The frontend and backend were seamlessly connected to ensure smooth communication between the user interface and the deepfake detection engine. The system was deployed on a cloud-based platform to ensure scalability and accessibility. Continuous monitoring and updates were implemented to maintain the system's performance and adapt to new challenges in deepfake technology. This end-to-end approach ensured that the system was not only functional but also future-proof, capable of evolving as new techniques and datasets became available.

F. User Experience and Feedback

To ensure the system was intuitive and accessible, significant emphasis was placed on user experience (UX) design. The interface was designed with simplicity and clarity in mind, allowing users to easily upload images or use real-time camera integration for deepfake detection. Feedback mechanisms were incorporated to gather user input, which was used to further refine the system. For instance, users could report false positives or negatives, providing valuable data to retrain and improve the model. Additionally, a comprehensive help section and tooltips were included to guide users through the process.

IV. RESULTS AND ANALYSIS

The AuthenticVision web application developed in this study addresses critical challenges in digital content authentication, focusing on enhancing user experience through integrated modules.



File uploaded successfully!

elon.jpg

Prediction: Real

fake2.jpg

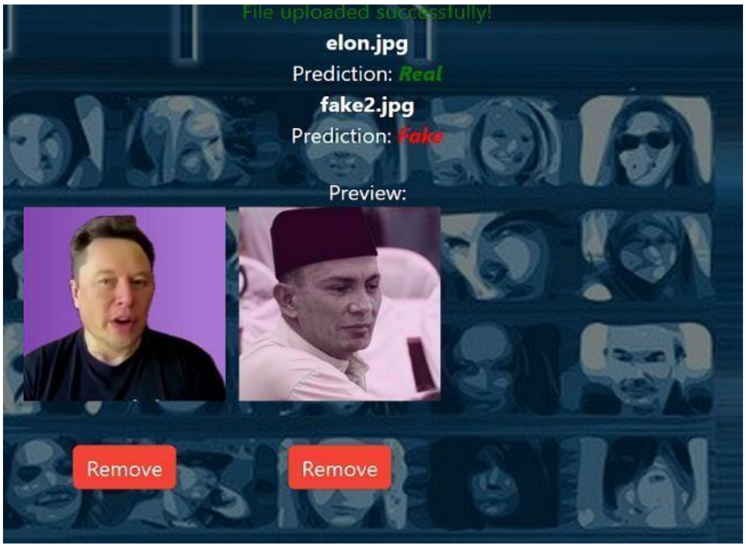
Prediction:

Preview:

Fig.2: Image Upload

Manipulation Explanation: This module simplifies complex technical findings for users without advanced expertise, providing clear, human-readable summaries of detected anomalies, such as pixel-level inconsistencies or semantic-level manipulations.

Manipulation Impact Assessment: The platform evaluates the potential impact of detected alterations, including trust erosion and misinformation propagation. Users receive actionable insights to prioritize critical manipulations and take necessary corrective actions.



File uploaded successfully!

elon.jpg

Prediction: Real

fake2.jpg

Prediction: Fake

Preview:

Fig.3: Detection

Performance and Usability: The system demonstrated robust performance under varying workloads, with low response times and reliable detection of manipulated content.

The convolutional neural network achieved a detection accuracy of 74%, with user testing confirming the platform's intuitive design and ability to meet diverse user needs effectively.

In conclusion, the application enhances trust, accessibility, and usability, empowering users to detect and understand deepfake content while improving overall digital content integrity through AI-driven solutions.

REFERENCES

1. Firmin, Sally, Kaur, Achhardeep, Noori Hoshyar, Azadeh, Saikrishna, *et al.*, "Deepfake video detection: challenges and opportunities." Springer Nature, 2024. [Online]. Available: https://core.ac.uk/download/619431140.pdf
2. IEEE Journals & Magazine, "Deep Learning Applications," Proc. of IEEE, vol. X, 2023. DOI: 10.1109/9839540.
3. Chen, Ziying. "Perception of crisis responsibility: examining ai-generated deepfake content and public response to taylor swift." Surface at Syracuse University, 2024.
4. IEEE Journals, "Image Segmentation Using Deep Learning: A Survey," Proc. of IEEE, vol. X, 2020. DOI: 10.1109/9356353.
5. Mahmud, Bahar Uddin, Sharmin, Afsana. "Deep Insights of Deepfake Technology: A Review." 2023. [Online]. Available: http://arxiv.org/abs/2105.00192
6. Bhattacharyya, Chaitali, Kim, Sungho, Wang, Hanxiao, Zhang, *et al.*, "Diffusion Deepfake." 2024.
7. Monteiro, Stéphane Mesquita. "Detection of fake images generated by deep learning." 2024. [Online]. Available: https://core.ac.uk/download /614512274.pdf