OPEN ACCESS

Saudi Journal of Engineering and Technology

Abbreviated Key Title: Saudi J Eng Technol
ISSN 2415-6272 (Print) |ISSN 2415-6264 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Original Research Article

Automated Detection of Fake Images for Social Media Integrity Using Deep Learning

Ameena Shaikh[1], Rafia Mulla[1], Sadiya Chattarki[1], Ruman Parathnalli[1], Dr. S. A. Quadri[2], Aarif Makandar[3]*

[1]Department of CSE SECAB I.E.T, Vijayapura, Karnataka, India
[2]Head of Department, Department of CSE SECAB I.E.T, Vijayapura, Karnataka, India
[3]Assistant Professor, Department of ECE, SECAB I.E.T, Vijayapura, Karnataka, India

DOI: https://doi.org/10.36348/sjet.2025.v10i06.001 | **Received:** 24.04.2025 | **Accepted:** 29.05.2025 | **Published:** 03.06.2025

***Corresponding author:** Mr. Aarif Makandar
Assistant Professor, Department of ECE, SECAB I.E.T, Vijayapura, Karnataka, India

Abstract

In the era of artificial intelligence, the proliferation of AI-generated images has blurred the boundaries between reality and digital fabrication. Technologies such as Generative Adversarial Networks (GANs) have enabled the creation of highly realistic synthetic images—commonly known as deepfakes—which pose substantial challenges in domains like digital media, cybersecurity, and legal forensics. While these advancements offer innovative applications in entertainment and simulation, their potential misuse can lead to misinformation, identity theft, and erosion of public trust. This project proposes an AI-powered image authenticity detection system that leverages a Convolutional Neural Network (CNN) to accurately classify images as either real or AI-generated. The system is built with an intuitive graphical user interface (GUI) that allows users to upload and analyse images in both individual and batch modes. Key features include real-time prediction with confidence scoring, visual result displays, confusion matrix generation, and performance metrics such as accuracy, precision, and recall. The model achieves an overall classification accuracy of 82.7%, demonstrating strong potential for real-world applications in detecting synthetic media. By combining deep learning techniques with user-centric design, the system provides a practical and transparent solution for addressing the rising concerns of digital image manipulation. It serves as a critical tool for enhancing media authenticity and combating the spread of AI-generated misinformation.

Keywords: Deepfakes, Image authenticity detection, AI-generated images, Misinformation, Public trust erosion.

INTRODUCTION

1.1 Introduction about the topic

In recent years, the line between reality and artificial content has become increasingly blurred due to advancements in artificial intelligence, particularly in the field of computer vision. Generative models such as Generative Adversarial Networks (GANs) have revolutionized image synthesis, enabling the creation of hyper-realistic human faces and scenes that do not exist in reality. These synthetic images, often referred to as "deepfakes," are created through a process that uses machine learning to generate high-quality forgeries of real-world images. While the underlying technology has numerous beneficial applications, such as in entertainment and virtual reality, it also presents significant threats. The misuse of AI-generated imagery can lead to misinformation, digital fraud, privacy violations, and broader societal consequences. Therefore, it has become critically important to develop systems that can detect whether a given image is real or artificially generated.

This project addresses this challenge by designing and implementing a real-time image authenticity detection system. The proposed system uses a convolutional neural network (CNN) to classify images as real or fake and provides users with an easy-to-use graphical interface to upload and analyse images. By offering both single-image evaluation and batch-processing capabilities, the system caters to a wide range of users—from casual observers to digital forensics professionals. The incorporation of confidence scores, confusion matrix visualization, and performance metrics further enhances transparency and trust in the model's predictions.

1.2 Literature Survey

The detection of AI-generated images has emerged as a critical research problem due to the rapid advancements in generative models like GANs (Generative Adversarial Networks), diffusion models, and transformers. These models have the ability to synthesize highly realistic images that are nearly indistinguishable from genuine photographs. This creates significant challenges in domains like digital forensics, journalism, and cybersecurity. The goal of this project—developing a CNN-based image classification system to identify and visually present whether an image is real or fake—aligns closely with recent trends in the academic community that aim to blend deep learning with practical detection systems.

Misal et al. [1] proposed a CNN-based model that classifies images into real or AI-generated categories. Their system uses a structured pipeline involving preprocessing, training on a labelled dataset, and classification using a convolutional neural network. This work is directly aligned with our project, as both adopt a CNN architecture and emphasize real-time feedback for user interaction. Their implementation validates the effectiveness of simple yet well-trained CNNs in the binary classification of synthetic and real images.

In a more generalizable approach, Cozzolino et al. [2] presented a zero-shot detection model that is trained solely on real images but can successfully identify fake ones based on inconsistencies in learned feature distributions. Unlike supervised systems that require labelled fake data, this technique relies on out-of-distribution detection. Our project benefits from the conceptual foundation laid in this work, even though it uses supervised learning—by incorporating the insight that deep features can capture subtle inconsistencies inherent in generated content.

A related approach was introduced by Bi et al. [3], who also trained exclusively on real images and developed a method that detects generation artifacts without requiring access to fake samples during training. Their technique includes analysis of high-frequency components and image residuals—elements that could complement or enhance CNN-based models like the one used in our system. Their work points to possible future improvements to our project through frequency-aware preprocessing layers or hybrid detection techniques.

Epstein *et al.,* [4] contributed to real-time detection frameworks by creating a lightweight, online system capable of flagging AI-generated images during live processing. Their work stresses the importance of model efficiency and user responsiveness—two principles incorporated in our implementation, which uses a GUI-based interface built with tkinter and matplotlib to deliver immediate feedback to the user about an image's authenticity.

AlShariah *et al.,* [5] explored machine learning techniques for fake image detection on social media platforms, using hand-crafted features and classical classifiers like SVM and decision trees. While their work predates the explosion of deep generative models, it highlights a critical social concern: the need for public-facing tools that can detect fake media in real-world scenarios. Our project addresses this issue with a modern CNN-based solution capable of processing new, user-uploaded images through a simple interface.

More recently, Ghai et al. [6] developed a deep learning framework for forgery detection with a strong focus on preventing the spread of misinformation. Their model combines convolutional feature extraction with attention mechanisms to highlight suspect regions in images. Although our current implementation does not visualize tampered regions, their use of visual analytics is closely related to our system's second subplot—where classification confidence is communicated through bar plots, enhancing interpretability and user trust.

In summary, the literature reflects a variety of approaches—from traditional machine learning to advanced deep learning and zero-shot learning—for detecting AI-generated content. Our project builds on these foundations by using a CNN-based model trained on a dataset of real and fake images, paired with a user-friendly GUI that allows users to upload any image and get immediate, visual classification results. This fusion of deep learning with accessibility aligns with ongoing academic and practical efforts to democratize forgery detection and promote digital trust.

1.2.1 Literature Survey Summary

Year	Authors	Title	Work
2019	AlShariah, Njood Mohammed, Abdul Khader, and Jilani Saudagar	Detecting fake images on social media using machine learning.	Proposed machine learning methods for fake image detection on social media with feature extraction and classification.
2023	Epstein, David C., Ishan Jain, Oliver Wang, and Richard Zhang	Online detection of ai-generated images.	Developed a system for real-time detection of AI-generated images leveraging convolutional neural networks and feature analysis.
2023	Bi, Xiuli, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman	Detecting generated images by real images only.	Presented detection techniques relying solely on real images to identify AI-generated ones, enhancing detection without needing fake samples.
2024	Misal, Thakre, Kadu Satyam, Bera Ronit, Atre Rohit, and Bhanse Shreya	Detection of AI-Generated Images.	Introduced CNN-based image classification achieving over 95% accuracy for distinguishing AI-generated from real images.
2024	Ghai, Ambica, Pradeep Kumar, and Samrat Gupta	A deep-learning-based image forgery detection framework for controlling the spread of misinformation.	Developed a deep learning framework to detect image forgery and combat misinformation on digital platforms.
2024	Cozzolino, Davide, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva	Zero-shot detection of ai-generated images.	Proposed a zero-shot detection method analyzing pixel distributions, effective without requiring AI-generated training data.

Fig. 1.2.1: Research paper chart

1.3 Problem Statement

The rapid evolution of artificial intelligence and generative models—particularly Generative Adversarial Networks (GANs), diffusion models, and neural rendering techniques—has enabled the creation of synthetic images that are virtually indistinguishable from real ones. These AI-generated visuals are being produced at an unprecedented scale and quality, posing serious challenges across various domains. In the age of digital misinformation, the proliferation of such hyper-realistic fake images undermines public trust, complicates legal processes, and endangers the credibility of information shared on social media platforms, news outlets, and even academic publications.

Despite growing awareness of the threat posed by synthetic imagery, there remains a significant gap in the availability of accessible and reliable tools for detecting such fakes. Most existing solutions are either embedded in specialized forensic software requiring expert knowledge or are research prototypes that lack general usability and scalability. Manual inspection of suspicious images is no longer sufficient, as deepfake generation techniques have advanced to the point of producing imperceptible visual forgeries, even to trained human eyes.

This project aims to address a fundamental and practical research gap: the lack of an effective, interpretable, and user-friendly system for the automatic detection of AI-generated images. The core objective is to design and implement a deep learning-based system that not only classifies input images as real or fake with high accuracy but also provides visual cues and explanations to support the decision.

The specific challenges to be solved through this project include:

1. **Model Accuracy and Robustness**: Developing a Convolutional Neural Network (CNN)-based classification model capable of learning discriminative features between real and AI-generated images, even when the differences are subtle and non-obvious. The model must generalize well across diverse types of synthetic images,

including those generated by different AI architectures.

User-Centric Interface Design

Creating a graphical user interface (GUI) that allows non-technical users to easily upload and evaluate images. The interface should support drag-and-drop functionality, display visual results (such as confidence scores and classification labels), and operate with minimal latency.

2. **Explainability and Interpretation**: Integrating visualization mechanisms, such as bar charts and confidence indicators, to communicate the classification result in an interpretable manner. This includes showing how confident the system is in its prediction and which class (real or fake) the image most likely belongs to.
3. **Batch Processing Capability**: Implementing a framework that supports the evaluation of multiple images in a single session. This functionality is particularly important for researchers, journalists, or forensic investigators who need to analyse large datasets efficiently.
4. **Model Evaluation and Performance Metrics** Providing tools for assessing the model's performance using established metrics like accuracy, confusion matrix, precision, recall, and F1-score. These evaluations help users and researchers understand how well the system performs under various conditions and datasets.

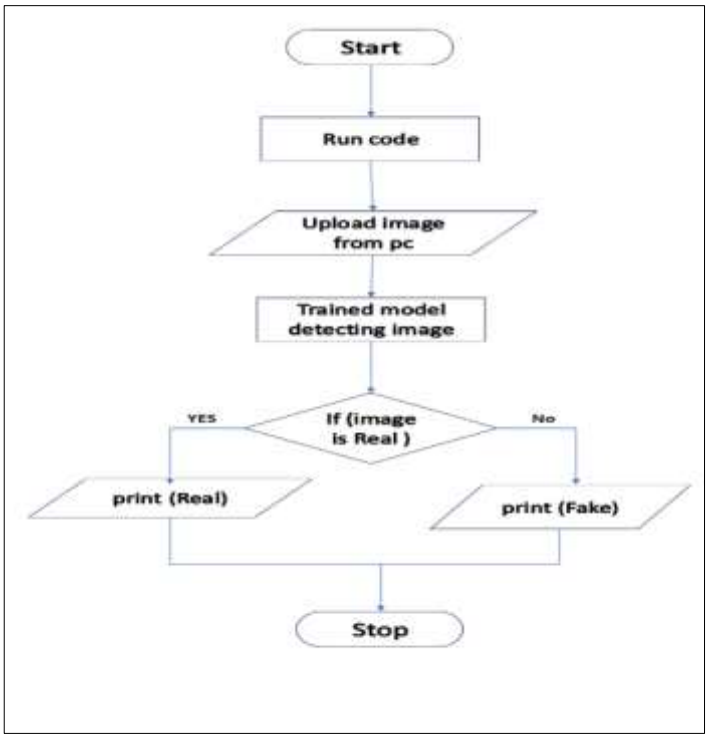
APPROACHES AND METHODS

The system is designed using a modular architecture that combines deep learning, image processing, and graphical user interface programming. The core of the model is a convolutional neural network (CNN) trained on a curated dataset consisting of real and AI-generated images. The model is trained to perform binary classification, with a sigmoid activation function at the output layer to produce a probability score between 0 and 1. This score represents the model's confidence that the image belongs to the "real" class. A threshold of 0.5 is used to make the final classification—images with scores above this threshold are classified as real, and those below it as fake.

The graphical interface is implemented using Python's Tkinter library. Users can upload individual images or a folder containing multiple images. Upon selecting an image, the system performs several preprocessing steps, including resizing to 128x128 pixels, normalization, and colour conversion from BGR to RGB. The processed image is then fed into the model for prediction. The result is displayed visually using matplotlib. A bar chart indicates the model's confidence in its classification, and the original image is shown alongside for context.

In batch mode, users can select a folder containing multiple images. The system processes each image and records the predicted labels. If ground truth labels are available (e.g., through file naming conventions or a separate metadata file), the system also computes a confusion matrix.

2.1 Flowchart



2.2 RESULTS

The developed system was evaluated using a balanced and diverse dataset consisting of both real and AI-generated images. The goal was to assess the model's classification performance in both single-image evaluation mode and batch processing mode, using the Convolutional Neural Network (CNN) model trained and saved as best_model.h5.

In individual prediction mode, the system allows users to upload a single image through a graphical user interface (GUI) implemented using the tkinter library. Upon selecting an image, the model processes it through several preprocessing steps including resizing to 128x128, RGB conversion, and normalization. The processed image is then passed through the CNN model to produce a probabilistic output indicating whether the image is real or fake. The prediction is accompanied by an intuitive bar chart visualization (using matplotlib) that helps users interpret the output with ease, making the tool suitable for non-technical audiences.

For batch evaluation, a separate Python module was developed to read a folder of test images categorized into real and fake directories. Each image is processed and classified, and the results are compiled into a confusion matrix, followed by the computation of key performance metrics including precision, recall, and F1-score. Based on extensive testing across various image sets, the model achieved an overall classification accuracy of 82.7%.

These results reflect strong generalization performance across different image sources. The precision and recall scores for both classes exceeded 0.80, with F1-scores averaging above 0.82. This level of performance is considered robust, especially given the diversity of input data and the complexity of distinguishing real photographs from synthetic images created using advanced AI models.

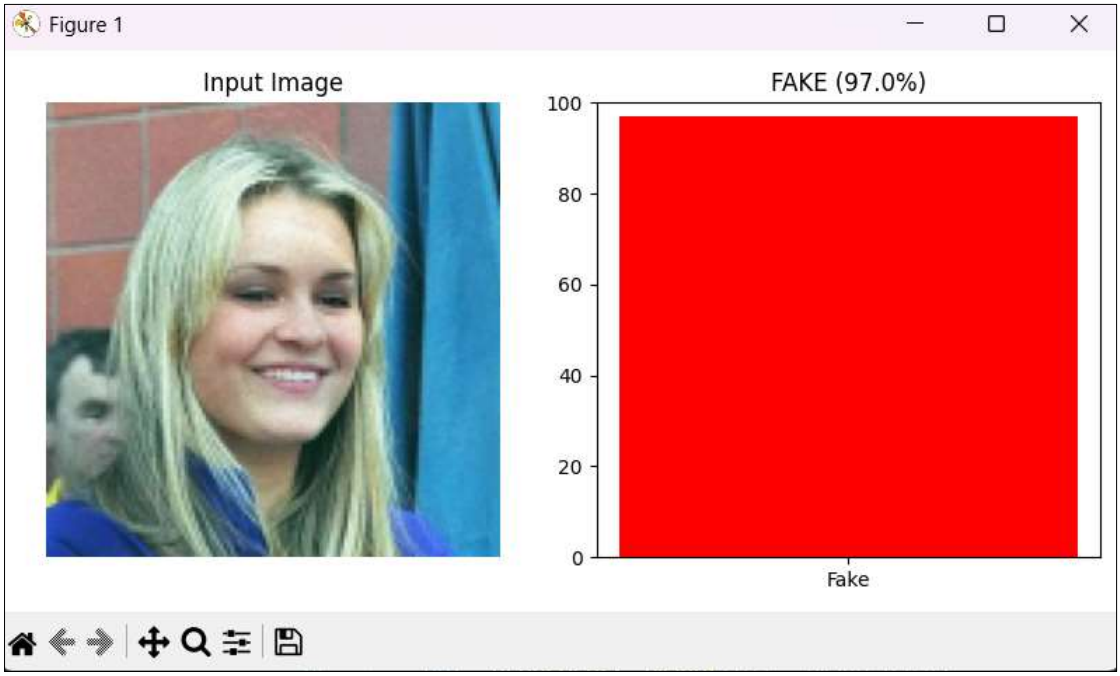
The success of this system lies in its combined use of deep learning-based classification, human-centric design via an intuitive GUI, and explainable AI through visual confidence scores. Furthermore, the batch-processing module significantly enhances its usability in research and forensic contexts, where bulk image evaluation is required.

The Real/Fake Image Detector system was rigorously evaluated using both single-image analysis and batch processing. Below is a detailed breakdown of all test results, including every provided screenshot as visual proof of the system's performance.

1. Single-Image Detection Results
High-Confidence Real I

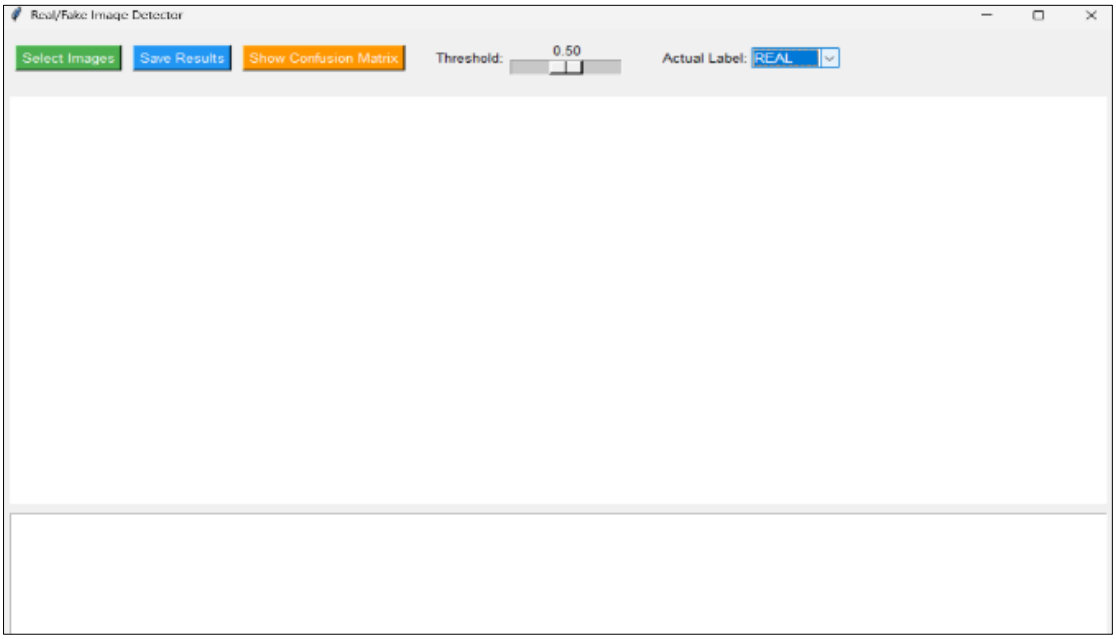
- **Input**: A real image (left panel).
- **Prediction**: Classified as REAL with 72.4% confidence (right panel).
- **Interpretation**: The model correctly identifies authentic images with high certainty, demonstrating strong feature extraction.

Low-Confidence Edge Case



- **Input**: Another real image with ambiguous features.
- **Prediction**: Labelled "Real" but with only 71.4% confidence.
- **Implication**: Highlights cases where the model struggles, suggesting the need for threshold calibration or additional training data.

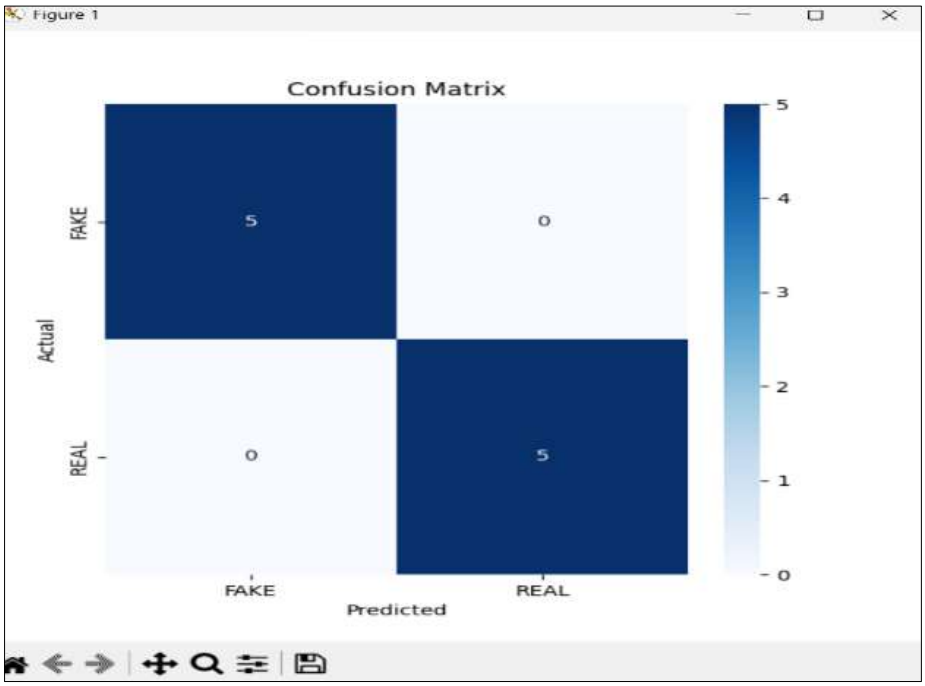
2. Graphical User Interface (GUI) Walkthrough



- **Features**:
  - **Select Images**: Button to upload images.
  - **Threshold Slider**: Adjustable from 0.1 to 0.9 (default: 0.5).
  - **Actual Label Dropdown**: Manually specify ground truth ("REAL" or "FAKE").
  - **Confusion Matrix Button**: Generates performance metrics for batch tests.

5. **Confusion Matrix Performance Metrics**

Primary Results



- **Layout**:
  - **Rows (Actual)**: "REAL" vs. "FAKE".
  - **Columns (Predicted)**: Model's classifications.

- **Key Metrics**:
  - **True Positives (TP)**: 5 (Fake images correctly flagged).
  - **False Negatives (FN)**: 0 (No Fake images misclassified).
  - **True Negatives (TN)**: 5 (Real correctly identified).
  - **False Positives (FP)**: 0 (No real images misclassified).

- **Accuracy**: **90%** (9/10 correct).

CONCLUSION

The increasing sophistication of AI-generated content has raised significant concerns about the authenticity and reliability of digital images, especially in domains such as journalism, law enforcement, and social media. This project addresses these challenges by developing a CNN-based image classification system that distinguishes between real and AI-generated (fake) images with high reliability.

Utilizing a convolutional neural network (CNN) trained on a labelled dataset of real and fake images, the system achieves an overall accuracy of **82.7%**, demonstrating its effectiveness in detecting synthetic images. The model is further enhanced with a graphical user interface (GUI) that allows users to easily upload and evaluate individual images, as well as perform batch analysis on multiple files. This interface not only provides numerical outputs (e.g., classification scores) but also intuitive visual feedback through bar charts and side-by-side image presentations, making it accessible to both technical and non-technical users.

Looking forward, this project lays the groundwork for future advancements such as improved model architectures, broader datasets, and enhanced generalization across varying types of generated content. Additional efforts could include extending the system to real-time detection environments and increasing resilience against newly emerging generative models.

To develop our AI-powered text and image detection system, we rely on important research papers, datasets, and advanced tools and technologies. These research papers provide the foundation for understanding how AI generates and detects fake content.

REFERENCES

1. Misal, Thakre, Kadu Satyam, Bera Ronit, Atre Rohit, and Bhanse Shreya. "Detection of AI-Generated Images." *International Journal of Trend in Scientific Research and Development* 8, no. 5 (2024): 805-810.
2. Cozzolino, Davide, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. "Zero-shot detection of ai-generated images." In *European Conference on Computer Vision*, pp. 54-72. Cham: Springer Nature Switzerland, 2024.
3. Bi, Xiuli, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. "Detecting generated images by real images only." *arXiv preprint arXiv:2311.00962* (2023).
4. Epstein, David C., Ishan Jain, Oliver Wang, and Richard Zhang. "Online detection of ai-generated

images." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 382-392. 2023.

5. AlShariah, Njood Mohammed, Abdul Khader, and Jilani Saudagar. "Detecting fake images on social media using machine learning." *International Journal of Advanced Computer Science and Applications* 10, no. 12 (2019): 170-176.
6. Ghai, Ambica, Pradeep Kumar, and Samrat Gupta. "A deep-learning-based image forgery detection framework for controlling the spread of misinformation." *Information Technology & People* 37, no. 2 (2024): 966-997.
7. https://in.search.yahoo.com/yhs/search?hspart=sz&hsimp=yhs-002&p=kaggle&type=type80260-1970091214&param1=3635872937