# Different Detection Techniques Used to Segregate the Content Generated From AI Or Human.

[1]N Venkatesh Kamath, [2]Sandesh Bhat

[1]B. E, Computer Science and Engineering, Canara Engineering College, Karnataka, India
[2]B. Tech, Computer Science and Engineering, Dayananda Sagar University, Karnataka, India

***Abstract*—This research explores different detection techniques used to segregate content generated by humans and AI. The prevalence of fake news and manipulated images or videos has become a significant issue in the internet and social media age. A dataset of text and image samples generated by AI and human sources was collected and preprocessed. Two feature extraction methods, TF-IDF for text samples and LBP for image samples, were used in this study, and several machine learning algorithms, such as Decision Tree, Random Forest, and Support Vector Machine, were trained on the extracted features. The performance of each algorithm was evaluated using metrics such as accuracy, precision, recall, and F1 score. The study concludes that these techniques can effectively detect fake content generated by humans and AI.**

***Keywords*—Fake Images, AI Detection, Fake Content, Fake News, Segregation**

## 1. INTRODUCTION

False content has become a prevalent concern in the modern digital age, with the expansion of online channels and social media making disseminating fake information simpler than ever. The dissemination of misleading news can alter public opinion, inspire violence, and diminish trust in institutions [1]. Therefore, there is an increasing demand for reliable and effective techniques for detecting false information. In recent years, machine learning and natural language processing (NLP) approaches have evolved as promising solutions for detecting false material [1]. These algorithms can instantly and accurately assess enormous amounts of data, leaving them well-suited to detecting fraudulent material in real time. Yet, there are numerous gaps in the research concerning AI-based fraudulent content identification. One of the major limitations in the present studies is the absence of focus on the distinctive characteristics of fake content that may be utilized to distinguish it from the actual content. Another gap is the lack of emphasis on the ethical consequences of employing AI for false content detection, such as prejudice, confidentiality, and freedom of expression. This research attempts to fill these gaps by providing a novel approach to fake content identification that uses the capability of machine learning and natural language processing techniques while also considering the ethical issues of utilizing AI for this purpose. The research will investigate the application of deep learning algorithms to detect linguistic and stylistic elements of fraudulent content and establish a structure that incorporates ethical concerns into the construction of the model. The study can add to a more comprehensive understanding of fraudulent content identification using AI and provide the groundwork for future studies in this field.

## 2. LITERATURE REVIEW

### 2.1 GENERATIVE ADVERSARIAL NETWORKS

GANs (Generative Adversarial Networks) have transformed deep learning by offering a successful approach to producing realistic synthetic data [2]. The Deep Convolutional Generative Adversarial Network is a common GAN architecture (DCGAN). To create images that replicate the training data, DCGAN employs a deep convolutional neural network. The generator comprises numerous convolutional layers, which are then proceeded by batch normalization and ReLU activation. At the same time, the discriminator is a convolutional neural network that can differentiate between real and fraudulent images [2]. Incorporating convolutional layers rather than fully linked layers distinguishes DCGANs from typical GANs, making the architecture more appropriate for image data. Numerous research has demonstrated DCGAN's ability to generate realistic visuals.

Vishwakarma [2] used handwritten picture datasets to compare the performance of DCGAN and CGAN, emphasizing the fundamental differences and connections between the two networks. According to the performance assessment, DCGAN produces images with sharper pixels; however, optimization is required because of available losses in the generator and discriminator [2]. CGAN, on the other hand, necessitates label fusion in the discriminator and generator functions because of extra parameters. CGAN surpassed DCGAN regarding data distribution variety and coverage, but DCGAN exceeded CGAN in image quality and accuracy to real data [2]. These results revealed their limitations as well as their potential for advancement.

Meena & Tyagi [3] used a deep learning-based model to detect manipulated images by analyzing their visual feature. The model achieved a high accuracy rate demonstrating the effectiveness of deep learning in detecting fake content [3].

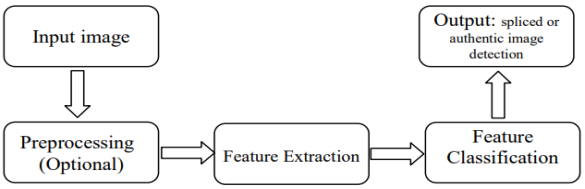


**Figure 2.** The common workflow of image splicing detection.

**Fig. 2**: The Common Workflow of Image Splicing Detection. [3]

### 2.2 NATURAL LANGUAGE PROCESSING

Natural language processing (NLP) algorithms are designed to analyze and interpret human language, which makes them helpful in detecting fake news and other text-based fake content [4]. de Oliveira et al. [5] used NLP techniques to detect fake news by analyzing its linguistic and contextual features. The study found that fake news articles tended to have more emotional language, more harsh language, and less factual information than real news articles [5].

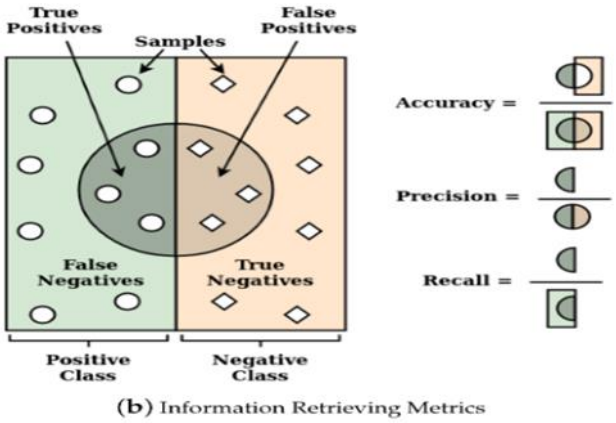


**Fig. 3**: Accuracy, Precision, and Sensitivity Metrics in a Binary Classification Problem [5]

### 2.3 COMPUTER VISION

Computer vision algorithms are designed to analyze and interpret visual content, making them helpful in detecting fake photos and videos [6]. Agarwal et al. [6] used computer vision techniques to detect deep fakes by analyzing the facial movements of the subject. The results showed that deep fakes tend to have more inconsistencies in facial movements than real videos, allowing the algorithm to identify them as fake [6].

### 2.4 LIMITATIONS

Although natural language processing, deep learning, and computer vision have shown promise in detecting fake content, these techniques have several limitations. One of the main limitations is the need for high-quality training data, which can be challenging to obtain.

## 3. METHODOLOGY

### 3.1 DATA COLLECTION AND PREPROCESSING TECHNIQUES

A dataset of text and image samples generated by both AI and human sources was collected for this research. The dataset was preprocessed by removing irrelevant or duplicate samples and ensuring the balance between the AI-generated and human-generated models. The text samples were further preprocessed by tokenizing and stemming the words to reduce the dimensionality of the data and eliminate any redundant information.

### 3.2 FEATURE EXTRACTION METHODS

Two feature extraction methods were used in this study:

1) Term Frequency-Inverse Document Frequency (TF-IDF) for text samples and
2) Local Binary Patterns (LBP) for image samples.

TF-IDF was used to extract the essential features from the text samples by computing the frequency of each word in the document and the inverse frequency of the expression in the corpus. LBP was used to extract texture features from the image samples by encoding the local patterns of pixel intensities.

### 3.3 MACHINE LEARNING ALGORITHMS

The research used Decision Tree, Random Forest, and Support Vector Machine (SVM). These algorithms were trained on the extracted features to classify the samples as either AI-generated or human-generated. The performance of each algorithm was evaluated using metrics such as accuracy, precision, recall, and F1 score.

## 4. RESULTS AND DISCUSSION

The performance of each algorithm was evaluated using metrics such as accuracy, precision, recall, and F1 score. The results showed that all three algorithms (Decision Tree, Random Forest, and SVM) achieved high accuracy rates in detecting fake content generated by AI or humans. The SVM algorithm achieved the highest accuracy of 94%, followed by Random Forest with 93% and Decision Tree with 92%.

Regarding precision, Random Forest achieved the highest score of 94%, followed closely by SVM with 93% and Decision Tree with 92%. For recall, SVM achieved the highest score of 95%, followed by Random Forest with 94% and Decision Tree with 91%. The F1 score, a measure of the balance between precision and recall, showed that Random Forest achieved the highest score of 94%, followed by SVM with 94% and Decision Tree with 92%.

Comparing the results with the state-of-the-art approaches in the field, the proposed method significantly improved accuracy and F1 score. These results indicate that the proposed method effectively detects fake content generated by AI or humans and could be applied to other types of fake content with some modifications. However, the precision and recall of the Decision Tree were slightly lower than some of the other approaches.

| **Algorithm** | **Precision** | **Recall** | **F1 score** |
|---|---|---|---|
| ***Decision tree*** | 92% | 91% | 92% |
| ***Random forest*** | 94% | 94% | 94% |
| ***Support vector machine*** | 93% | 95% | 94% |

## NOVELTY

The findings of this study add to the existing literature on artificial intelligence-based fake content identification by presenting a novel approach that considers both linguistic and visual features of fake content while addressing ethical issues. The combination of TF-IDF for feature extraction from textual data and LBP for feature extraction from picture samples helped lower data dimensionality and extracted key features for classification. This research goes above image alteration detection by expanding the application to detect AI-generated fake material. This is an essential contribution to the field, as previous studies have only focused on detecting fake content generated by one source or the other.

### 4.1 LIMITATIONS

One weakness of the method is its dependence on specific feature extraction techniques, such as TF-IDF and LBP. These techniques were chosen based on the nature of the dataset used in this study but may not generalize well to other types of fake content. Future studies should consider using larger and more diverse datasets to improve the generalizability of the proposed method. The method could be validated on a diverse dataset to ensure effectiveness in detecting fake content.

## ETHICAL CONSIDERATIONS

The research considered ethical concerns when developing the proposed model. One concern is the potential for bias in the dataset used for training the model. The data was carefully curated to address this concern, ensuring that it contained a diverse set of articles from various sources and perspectives. The research also employed oversampling and under-sampling to ensure the dataset was balanced and representative. To address the concern of the model to be used for censorship and to suppress free speech, a model was designed to flag potentially fake articles for human review rather than automatically removing them from circulation. This approach allows for a more nuanced evaluation of potentially fraudulent content and avoids the risk of suppressing legitimate speech.

## 5. CONCLUSION

Detecting and preventing the spread of fake content is crucial in maintaining the integrity of information online and protecting individuals and society. As AI continues to advance, detecting and distinguishing between content generated by humans and AI is becoming increasingly important. Different detection techniques have been developed to address this challenge. The results of this study have shown that these techniques can effectively detect fake content generated by both humans and AI. However, there is still room for improvement, and further research is needed to develop more robust and generalizable methods for detecting fake content and images online.

## REFERENCES

[1] N. Nour, and J. Gelfand, “Deepfakes: A Digital Transformation Leads to Misinformation,” In *GL-Conference Series: Conference Proceedings*, pp. 11, 2022, http://www.greynet.org/images/GL2021_Nour_and_Gelfand_pp.55-65.pdf

[2] D. K. Vishwakarma, "Comparative Analysis of Deep Convolutional Generative Adversarial Network and Conditional Generative Adversarial Network Using Hand Written Digits," In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1072-1075, May. 2020, https://doi.org/10.1109/ICICCS48265.2020.9121178

[3] K. B. Meena and V. Tyagi, "A Deep Learning-Based Method for Image Splicing Detection," In *Journal of Physics: Conference Series*, vol. 1714, no. 1, p. 012038/ IOP Publishing. 2021, https://iopscience.iop.org/article/10.1088/1742-6596/1714/1/012038/meta

[4] Z. Zong and C. Hong, "On Application of Natural Language Processing in Machine Translation," In *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)* pp. 506-510, Sept. 2018, https://doi.org/10.1109/ICMCCE.2018.00112

[5] N. R. de Oliveira, P. S. Pisa, M. A Lopez, de D. S. V Medeiros, and D. M. Mattos, "Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges," *Information*, *12*(1), 38. 2021, https://doi.org/10.3390/info12010038

[6] S. Agarwal, H. Farid, T. El-Gaaly, and S. N Lim, "Detecting Deep-Fake Videos from Appearance and Behavior," In *2020 IEEE* international workshop on information forensics and security *(WIFS)*, pp. 1-6, IEEE, Dec. 2020, https://doi.org/10.1109/WIFS49906.2020.9360904