# INDIAN INSTITUTE OF INFORMATION TECHNOLOGY ALLAHABAD

# Identification of Artificially Generated Images

---

Group Member
Shubham Kumar Bhokta
IIT2020007

**Project Supervisor**
Dr. Shiv Ram Dubey

---

## Abstract:

In this paper, we present a novel approach for distinguishing between real and fake images using Deep Convolutional Neural Networks (DNNs). Specifically, we employ the ResNet architectural structure to address the challenge of increasingly convincing synthetic imagery in image processing and computer vision domains. Our methodology encompasses extensive data collection, preprocessing, feature extraction, model selection, training, and evaluation. We emphasize the significance of accuracy, recall, and precision metrics in assessing model performance. Through the implementation of ResNet50 and rigorous testing, our approach achieves a remarkable accuracy of 77%, a recall of 75%, and an F1 score of 74%. This framework contributes substantially to content authentication and safeguarding the integrity of visual information in the digital age. Furthermore, it offers valuable applications in combating disinformation and enhancing media credibility. As technological advancements progress, collaborative efforts and public awareness remain paramount in ensuring the authenticity and reliability of visual content.

**Keywords:** CNNs, DNNs, ResNet, ResNet50, accuracy, recall, precision, GANs, AI

## I. INTRODUCTION

In the fields of image processing and computer vision, the spread of artificial intelligence technology has led to the creation of convincingly deceptive artificially created images. These synthetic images, often created using deep learning-based models, challenge the authenticity and reliability of visual content. In this paper, we present a comprehensive approach to artificially generated image recognition using a ResNet-based framework.

We have used the ResNet model. The ResNet model, known for its exceptional feature extraction capabilities, enables us to construct an effective classifier that is capable of detecting the signs of artificial image generation.

Our results demonstrate the model and its ability to accurately distinguish between genuine and manipulated images, contributing to the critical task of content authentication and ensuring the integrity of visual information in the digital age. The proposed method has a wide range of applications, from combating disinformation and image forgery to improving the credibility of digital media. As image processing techniques become increasingly sophisticated, our ResNet-based solution provides a reliable way to distinguish between real and artificially created. It is a valuable tool to ensure the authenticity and credibility of visual content in various fields.

## II. Literature review

| S.No | *Authors* | *Paper title* | *Description* | *Methodology* | *Result* |
|---|---|---|---|---|---|
| 1. | Jordan J. Bird, Ahmad Lotfi | Image Classification and Explainable Identification of AI-Generated Synthetic Image | Recent advances in synthetic data technology enable the creation of highly realistic images that are indistinguishable from real photos. This article suggests utilizing computer vision to identify AI-generated images and creating a synthetic dataset similar to CIFAR-10 with latent diffusion for comparison. A Convolutional Neural Network (CNN) is then employed to classify images as real or AI-generated, achieving a 92.98% accuracy rate after thorough training and hyperparameter optimization. The study also employs Gradient Class Activation Mapping for explainable AI, revealing that the model relies on minor background imperfections rather than the actual subjects for classification. The CIFAKE dataset, developed for this study, is now publicly accessible for future research. | This study utilized synthetic data resembling CIFAR-10 to enable comparison of AI-generated images. A Convolutional Neural Network (CNN) was employed to classify images as real or AI-generated, achieving a 92.98% accuracy rate. Explainable AI using Gradient Class Activation Mapping revealed that small background imperfections were critical for classification. The resulting CIFAKE dataset is now available for future research. | After extensive tuning and training, the CNN achieves a 92.98% accuracy rate. |
| 2. | Brandon Khoo1\| Raphaël C.-W. Phan1,2 \| Chern-Hong Lim | Deepfake attribution: On the source identification ofartificially generated images | This article discusses the rapid advancements in synthetic media, also known as "deep fakes," focusing on their improved visual quality and the challenges they pose in distinguishing them from real images. It highlights the need for reliable detection methods and a shift in research towards attributing AI-generated images to their sources. The article also explores the ethical considerations and the potential for holding malicious users accountable while protecting intellectual | Data Collection: Assemble a diverse dataset of images, real and AI-generated. Model Development: Create a deep learning model for deepfake detection and attribution. Training and Evaluation: Train the model on the dataset and assess its performance using key metrics. Limitation Analysis: Examine model limitations and potential counter-forensic attacks. Research and Ethical Considerations: Identify research | The article discusses advances in deepfake technology, the need for reliable detection methods, and the shift towards attributing AI-generated images to their sources. It also mentions the limitations of image recognition methods, counter-forensic attacks, and the importance of reliable attribution for security, |

| S.No | *Authors* | *Paper title* | *Description* | *Methodology* | *Result* |
|---|---|---|---|---|---|
| | | | property in deepfake technology. | directions and address ethical concerns in deepfake technology usage. | privacy, and intellectual property protection. |
| 3. | Diego Gragnaniello, Francesco Marra, and Luisa Verdoliva | Detection of AI-Generated Synthetic Faces | Recent advances in AI-driven synthetic media creation, particularly in generating lifelike human faces, have raised concerns about media trustworthiness and the proliferation of fake identities online. Detecting synthetic faces amidst real ones is a pressing challenge. While the scientific community is actively researching this issue, a universal detector is yet to be established. This ongoing cat-and-mouse game involves continuously improving detectors to counter increasingly realistic synthetic face generators. This chapter explores effective techniques for detecting synthetic faces, discussing their rationale, real-world applications, and comparative accuracy and generalization abilities. | Collect diverse synthetic and real human face images. Preprocess and augment the dataset. Extract features using a deep CNN. Train a classification model and evaluate it. Compare with existing methods, analyze real-world use, and assess generalization | Recent advances in AI-based synthetic media, especially for human faces, raise trust and identity concerns. Detecting synthetic faces remains a challenge, with ongoing research in this cat-and-mouse game. This chapter explores techniques, applications, and accuracy in differentiating synthetic from real faces. |

## III. OBJECTIVE

**1. Preventing Misuse of Synthetic Content:**

- Mitigating the malicious use of synthetic content, such as deepfakes, to spread misinformation, manipulate public opinion, or impersonate individuals.
- Safeguarding the integrity of digital media by ensuring that users can distinguish between authentic and synthetic images.
- Protecting individuals and organizations from being deceived by fake or altered content.

**2. Safeguarding People's Security and Privacy:**

- Preventing the unauthorized use of synthetic images to create fake identities or impersonate individuals, which could lead to fraud, cyberbullying, or other forms of online harm.
- Safeguarding individuals' privacy by detecting the unauthorized use of their images in synthetic content without their consent.

- Ensuring that synthetic images are not used to manipulate or exploit vulnerable individuals, especially in the context of online dating or other sensitive interactions.

**3. Upholding Trust, Authenticity, and Credibility:**

- Maintaining the credibility of information and digital media by verifying the authenticity of images.
- Building trust in online platforms and social media by combating the spread of fake news and disinformation.
- Ensuring that users can rely on the accuracy and integrity of visual content, particularly in domains such as journalism, forensics, and legal proceedings.

## IV. Deep Convolutional Neural Network

Deep convolutional neural networks (CNNs) are a type of artificial neural network designed to process and analyze visual information, such as images and videos. They have revolutionized the field of computer vision and are now widely used in image processing, object detection, and facial recognition systems.Key Characteristics of a Deep CNN:

1. **Convolutional Layers:** These are the fundamental building blocks of CNNs. They use convolution operations to extract features and patterns from the input image.
2. **Pooling Layers:** Pooling layers reduce the dimensionality of the feature maps produced by the convolutional layers. They help to make the network more robust to noise and variations in the input image.
3. **Fully Connected Layers:** After a series of convolutional and pooling layers, there are typically one or more fully connected layers. These layers are responsible for making predictions based on the features extracted by the earlier layers.
4. **Activation Functions:** Non-linear activation functions, such as the Rectified Linear Unit (ReLU), are applied after the convolution and fully connected layers to add non-linearity to the model. This allows CNNs to learn complex relationships in the data.
5. **Depth:** CNNs are often referred to as "deep" because they have many layers, often with a large number of filters in each layer. This depth allows them to learn hierarchical features, working their way up from basic edges to complex object representations.
6. **Weight Sharing:** One of the key ideas behind CNNs is weight sharing. The same set of learnable filters is applied to different regions of the input image in the convolutional layers. This weight sharing allows the network to learn translation-invariant features.
7. **Dropout:** Dropout is a regularization technique often used in CNNs to prevent overfitting. It involves randomly "dropping out" neurons from the training process by setting their outputs to zero. This forces the network to learn more robust features.
8. **Batch Normalization:** Batch normalization is another technique used to accelerate and stabilize training. It improves gradient flow and increases training stability by normalizing the inputs to each layer in the network.

## V. Residual Neural Networks (ResNets)

1. Residual neural networks (ResNets) are a type of deep learning architecture designed to address the vanishing gradient problem in extremely deep networks.
2. ResNets were first introduced in 2015.
3. They have had a significant impact on the

fields of computer vision and deep learning.

**Key Characteristics and Components of ResNets:**

1. **Residual Block:**
    - The fundamental building block of a ResNet.
    - Consists of two or more convolutional layers with a shortcut connection.
    - The shortcut connection adds the input directly to the output of the convolutional layers, bypassing some of them.
2. **Skip Connections:**
    - Help avoid the vanishing gradient issue during training.
    - Facilitate gradient flow, allowing deeper networks to be trained more effectively.
3. **Identity Mapping:**
    - In some cases, no additional convolutional layers are added, and the input is added directly to the output.
    - This is known as identity mapping when the input and output dimensions are the same.
4. **Bottleneck Architectures:**
    - Often used in deep ResNets to reduce the number of parameters and computational load.
    - Utilizes 1x1 convolutions to reduce dimensionality before applying 3x3 convolutions.
5. **Architecture Depth:**
    - ResNets can have extremely deep architectures, with hundreds or even thousands of layers.
    - Deep networks are essential for capturing hierarchical features and producing state-of-the-art results in various computer vision tasks.
6. **Batch Normalization:**
    - Frequently used in ResNets to stabilize and accelerate training by normalizing the inputs to each layer.

## VI. METHODOLOGY

### 1. Data Preparation:

- The dataset is organized into training, validation, and test sets, each containing real and fake images.
- The distribution of images in each dataset split is checked using the `check_dist()` function to ensure balance.

### 2. Data Preprocessing:

- Images are resized to a uniform size of 224x224 pixels using `IMG_SIZE`.
- Data augmentation techniques like rescaling are applied using `ImageDataGenerator` to improve model generalization.

### 3. Model Building:

- A convolutional neural network (CNN) model is constructed using the transfer learning approach with the ResNet50 architecture as the base model.
- The base model layers are frozen, except for the last 150 layers, to prevent overfitting and retain pre-trained weights.
- Global average pooling, dropout, and dense layers are added to the model for feature extraction and classification.
- The model is compiled with the Adam optimizer, categorical cross-entropy loss function, and accuracy metric.

### 4. Model Training:

- The model is trained on the training data (`train_flow`) for 15 epochs, with validation on the validation data (`valid_flow`).
- ModelCheckpoint and EarlyStopping callbacks are utilized to save the best model

and prevent overfitting.

### 5. Model Evaluation:

- The trained model is evaluated on the test data (test_flow) using the evaluate() method to obtain the test loss and accuracy.
- Additional evaluation metrics such as confusion matrix and classification report are computed using scikit-learn functions.

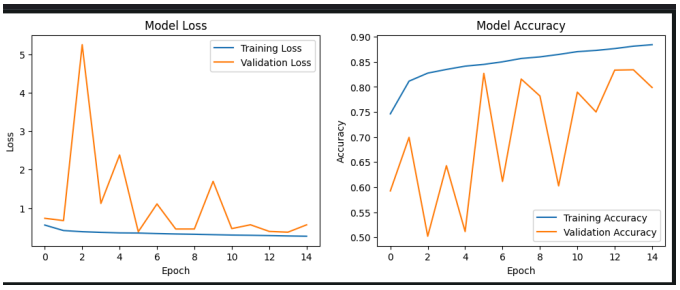
### 6. Reporting:

- Training history, including training/validation loss and accuracy over epochs, is visualized using Matplotlib.
- Individual image predictions are made and displayed alongside their actual and predicted labels for qualitative analysis.

### 7. Testing:

- Finally, we tested the fine-tuned model on the held-out testing set to assess its generalization performance.
- The model achieved an accuracy of 77%, a recall of 75%, and an F1 score of 74% on the testing set.

## VII. Results




Img 1. Model loss and accuracy

We have trained our model for 15 epoch, in which we got the following results while testing : The accuracy of our model is 77%, and recall is 75%, and f1 score is 74%.

```
10905/10905 [==============================] - 88s 8ms/step
Confusion Matrix:
[[3282 2210]
 [ 535 4878]]

Classification Report:
              precision    recall  f1-score   support

        Fake       0.86      0.60      0.71      5492
        Real       0.69      0.90      0.78      5413

    accuracy                           0.75     10905
   macro avg       0.77      0.75      0.74     10905
weighted avg       0.77      0.75      0.74     10905
```

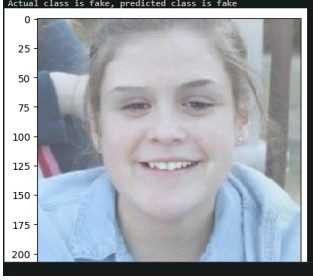
```
import matplotlib.pyplot as plt
img,label = test_flow[10];
label_ = label.argmax(axis=1)

res = model.predict(img)

class_ = res.argmax(axis=1)
if class_[0] == 0:
    if label_ == 0:
        print("Actual class is fake, predicted class is fake")
    else:
        print("Actual class is real, predicted class is fake")
else:
    if label_ == 0:
        print("Actual class is fake, predicted class is real")
    else:
        print("Actual class is real, predicted class is real")

plt.imshow(img[0])
plt.show()
```

```
1/1 [==============================] - 0s 26ms/step
Actual class is fake, predicted class is fake
```



## VIII Conclusions

In conclusion, identifying artificially generated images is increasingly critical in the digital age due to advancements in AI and deep learning, making it easier to create realistic fake images with potential for misuse. Researchers and technologists have developed innovative techniques, from traditional forensics to cutting-edge deep learning, to detect such images. Collaboration among experts in computer vision, machine learning, and digital forensics is essential. Educational efforts and public awareness are also vital in recognizing and countering fake imagery. As technology evolves, our detection and prevention strategies must adapt. Through vigilance, collaboration, and digital literacy, we can protect the authenticity and trustworthiness of the images that shape our world.

## IX. ACKNOWLEDGMENT

I am pleased to extend my heartfelt appreciation and gratitude to Dr. Shiv Ram Dubey, who served as our teacher and mentor throughout the duration of my project. Her invaluable guidance, support, and feedback played a crucial role in shaping my understanding of the subject matter and helped me achieve my goals.

I would also like to express my sincere thanks to the Teaching Assistants, for their exceptional support and assistance during the course of my project. Their expertise and dedication were instrumental in helping me develop my skills and achieve my objectives

## REFERENCES

[1] Bird, J. J., & Lotfi, A. (2023). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. arXiv preprint arXiv:2303.14126 [cs.CV]. Paper Link

[2] Khoo, B., Phan, R. C.-W., & Lim, C. H. (2021). Deepfake attribution: On the source identification of artificially generated images. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(7). DOI: 10.1002/widm.1438. Paper Link

[3] Gragnaniello, D., Marra, F., & Verdoliva, L. (2022). Detection of AI-Generated Synthetic Faces. In Handbook of Digital Face Manipulation and Detection (pp. 191–212). Advances in Computer Vision and Pattern Recognition ((ACVPR)). Open Access. Paper Link