Real VS AI Generated Image Detection and Classification

Mr. K. Balakrishna Maruthiram[1], Dr.G. Venkataramireddy[2], M. Klick[3]
[1]*Assistant Professor of CSE, Department of IT, JNTU Hyderabad, Hyderabad, India*
[2]*Professor of IT, Department of IT, JNTU Hyderabad, Hyderabad, India*
[3]*Student, M. Tech (Computer Networks and Information Security), Department of Information Technology, JNTU Hyderabad, Hyderabad, India*

***Abstract*-Recent advancements in synthetic data technology have made it possible to generate images of such exceptional quality that distinguishing between photographs taken in real life and those created by Artificial Intelligence has become nearly impossible for humans. Key features of system include single image processing, support for .jpg and .png file formats, and the provision of future scope for batch processing and download options. The classifier shows resilience and reliability in real-world scenarios through stringent testing and quality assurance procedures, such as unit testing, integration testing, and performance evaluation. Artificial intelligence can now create images of humans and other objects more easily because to the rapid advancement of deep learning technologies. One of the most important tasks in picture analysis and authentication is identifying manipulated or fraudulent images produced by artificial intelligence. It is critical to develop an accurate approach to identify these fake photographs. The goal of this project is to categorize photos into two groups: Artificial Intelligence-generated or real images using a Convolutional Neural Network (CNN).**

I. INTRODUCTION

In recent years, the field if AI-driven synthetic image generation has advanced rapidly, making the accurate detection of AI-generated photos essential to verify the authenticity if image data. Recently, generative technology has frequently resulted in images that have obvious aesthetic defects. But these days, AI models can produce realistic, high-fidelity photos in a matter of seconds. These artificial intelligence (AI)-generated images have improved to the point where they can now compete with real painters and even win art contests.

Image classification in pivotal in computer vision, bearing significant importance in both professional setting and daily life. This process encompasses image preprocessing, segmentation, extraction of key features, and identification through matching. Leveraging cutting-edge techniques in image classification enables quicker access to picture information, facilitating its application in scientific experiments with unprecedented efficiency.

In recent decades, the widespread adoption of social networks has deeply engaged people worldwide. Microblogging platforms have enabled individuals to share their thoughts in real-time on a global scale, providing researchers with valuable insights into online social dynamics during various events. This freedom of expression has facilitated the exchange of diverse thoughts, emotions, and knowledge among users. However, the digital environment isn't always secure, often becoming a platform for the dissemination of harmful content. Hate speech, a prevalent form of online expression, frequently manifests as prejudice, aggression, racism and other forms of verbal abuse.

II. EXISTING SYSTEM

Researchers have been looking at phony photos on social networking sites like Facebook, Instagram, and Twitter in recent years. Although there exist picture categorization techniques like Random Forest algorithms and Support Vector Machine, their accuracy has been restricted. Image processing and classification using these techniques is laborious and time-consuming. A number of critical parameters need to be set correctly in order to get the best categorization results.

It is crucial to distinguish between images produced by machine learning models and those that are real for a

number of reasons. Verifying the legitimacy and uniqueness of an image is essential. For example, an advanced Stable Diffusion Model (SDM) might produce a fake photo showing someone breaking the law, which could give false alibi proof for someone who was truly somewhere else. Today's environment is rife with fake news and disinformation, which can be used to influence public opinion through the use of machine-generated pictures.

III. PROPOSED SYSTEM

We proposed an algorithm that classify images into fake and AI-Generated images. This project focuses on developing a Convolutional Neural Network (CNN) model tailored for image classification, with specific emphasis on distinguishing between real and AI-generated images. Key components of the project include training the model using a curated dataset, creating a user interface for interaction, and compiling a detailed project report. The geographical scope is global, as the dataset comprises images from diverse regions. The project does not include the development of a mobile application or integration with external hardware systems. The model will be implemented using Python and TensorFlow, and the compatibility will be ensured with mainstream operating systems. The temporal scope spans six months. Data privacy and security measures will be implemented to protect sensitive information within the dataset. The project assumes the availability of computing resources and adherence to ethical standards in data collection and usage. In recent years, a great deal of research has been done to create automatic techniques for identifying AI-generated photos on social media or elsewhere. The task typically involves classifying images into Ai-generated or not. This is where we may utilize Image detection to identify whether the image is Ai-generated or not and can take relevant actions against the people who are spreading fake images and helps in preventing cyber-hate.

IV. ALGORITHM DESCRIPTION

1.Data Collection and Preprocessing

Dataset: This project uses a dataset consisting both real and Ai-generated images. Real images are sourced from various public image datasets, while AI-generated images can be collected from platforms like This Person Does Not Exist, GAN-generated datasets, etc.

Preprocessing: Images are resized to a standard dimension (e.g., 224*224 pixels) and normalized to a range of [0,1]. To improve model robustness, data augmentation techniques (such as flips, rotations, and color modifications) are used.

Convolutional Neural Networks (CNNs) are a basic type of deep learning models that are widely applied to computer vision applications like facial recognition, object identification, picture recognition, and classification. CNNs are made up of learnable parameters like weights and biases, just like simple neural networks. When processing images, which are represented as matrices of pixels (height by breadth by depth, or N*N*3), they are especially well suited.

A CNN comprises five essential components:
1. Input Layer
2. Convolution Layer
3. Pooling Layer
4. Fully Connected Layer
5. Output Layer

1.Input Layer

The input layer of a convolutional neural network (CNN) accepts raw image data. It transforms this data into a format suitable for processing by subsequent layers. This layer sets the stage for feature extraction.

2.The Convolution Layer

It makes use of a collection of learnable filters, each intended to identify particular patterns or features in the input image (usually indicated as M*M*3). Activation maps are produced by convolving (sliding) these filters across the input image's width and height and computing dot products. CNNs are invaluable in contemporary computer vision applications because of this mechanism, which aids in their ability to acquire hierarchical representations of visual data.

3.The Pooling Layer

In the CNN architecture, the pooling layer is situated in between the convolutional layers. Its main purpose is to reduce the network's computational load and parameter count. Pooling works by keeping the image's depth while decreasing its spatial dimensions. Generally, pooling is implemented separately for every depth dimension. Max pooling is a popular technique that retrieves the maximum value from each area of the picture that the pooling kernel covers.

4.Fully Connected Layer
High-level features are mapped to the final output via the fully linked layer. It guarantees that the learnt features from earlier layers serve as the foundation for the network's predictions. Feature maps are transformed into class scores or other outputs by this layer.

5.The Output Layer
Following several convolution and pooling layers, the output must be formatted into several classes. Key components in feature extraction and parameter reduction from original pictures are convolution and pooling layers. However, a fully connected layer is required to provide outputs corresponding to the number of desired classes in order to obtain the final classification output. Convolution layers generate 3D activation maps, however the resultant output for image categorization is usually either binary or categorical. To evaluate prediction accuracy, the output layer uses a loss function similar to categorical cross-entropy.

V IMPLEMENTATION

Import libraries: Importing Libraries launch by importing the necessary Python libraries for data running, visualization, and machine literacy tasks. Common libraries include NumPy, Pandas, and Matplotlib.

Import Dataset: Importing artificial intelligence (AI)-generated synthetic images from the CIFAKE dataset. To facilitate efficient model training and evaluation, the dataset will be suitably divided into training, validation, and testing sets. Extract the CIFAKE dataset into a working directory after downloading it from Kaggle. Keep distinct folders for photos produced by AI and those that are real.

Exploratory EDA: EDA (Exploratory Data Analysis) Perform exploratory data analysis is a crucial step in understanding the dataset and preparing it for model training. For the “Real vs AI Image Classifier”, EDA will help in gaining insights into the dataset, checking for class imbalances, and preprocessing the images.

Resize the images to fixed size (e.g., 224*224 pixels) to match the input size required by the pretrained models. Scale the pixel values to suitable range, typically between 0 and 1, to facilitate faster convergence during training.

Implying Algorithms: Using Keras to apply Convolution Neural Networks (CNN). Three convolution layers make up the model, and after each are max-pooling layers and dense layers for classification. We use the Adam optimizer and sparse categorical cross-entropy loss to construct the model. Image Data Generator is used to manage the augmentation and normalization of the preprocessed data. To track the model's performance, it is trained using the training data for a predetermined amount of epochs and then assessed using the validation set. Effective classification of actual vs. AI-generated photos is ensured by this method.

VI. OUTPUT AND RESULT

Precision: For class 0 (real images), precision is 0.88, indicating that 88% of predicted real images were actually real. For class 1 (AI-generated images), precision is 0.96, meaning 96% of predicted AI-generated images were correctly identified.

Recall: For class 0, recall is 0.96, indicating that 96% of actual real images were correctly predicted as real. For class 1, recall is 0.87, meaning 87% of actual AI-generated images were correctly identified as such.

F1-Score: The F1-score, which combines precision and recall into a single metric, is 0.92 for class 0 and 0.91 for class 1.

Support: Indicates the number of samples for each class (10000 for both class 0 and class 1).
The overall accuracy of the model is 91%, as calculated from the correctly predicted samples out of the total 20000 samples in the dataset. The macro average F1-score, precision, and recall are all 0.91, while the weighted average values are also 0.91, reflecting a balanced performance across both classes.

Classification Report :

	precision	recall	f1-score	support
0	0.88	0.96	0.92	10000
1	0.96	0.87	0.91	10000
accuracy			0.91	20000
macro avg	0.92	0.91	0.91	20000
weighted avg	0.92	0.91	0.91	20000

Fig.1 Classification Report

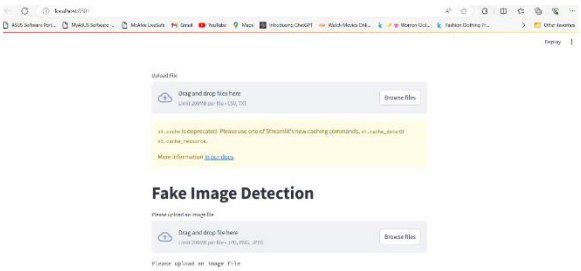

Fig.2 User Interface

The below figure shows the user interface of the website. In our project on classifying real vs. AI-generated images using a Convolutional Neural Network (CNN), we utilized Stream lit to create an interactive and user-friendly web application. Stream lit facilitated the rapid development of our application's front end, allowing us to easily deploy our trained model for real-time image classification. With Stream lit, we could effortlessly display images, show classification results, and provide an intuitive interface for users to upload their own images for analysis. This streamlined approach significantly enhanced the accessibility and usability of our machine learning model. In the website we have option to upload a jpg or png photos. We have an option browse files so that we can upload the images as per requirement. After uploading the image, we can get the image is real or fake i.e. AI-Generated Image. This user interface is created by using Python Library Stream Lit.

VII. CONCLUSION

In this project, we explored the critical task of distinguishing between real and AI-generated images using Convolutional Neural Networks (CNNs). The ability of AI to create highly realistic images has raised significant concerns regarding image authenticity and integrity. Our approach involved training a CNN on a dataset containing both real-life photographs and AI-generated images, leveraging advancements in deep learning and synthetic data generation technologies. Looking forward, the future of detecting AI-generated versus real images using Convolutional Neural Networks (CNNs) holds significant potential for advancement. Key areas for future development include: Ethical and Transparency Considerations: Developing frameworks that prioritize ethical considerations in the deployment of image classification technologies, including transparency in decision-making processes and adherence to privacy and fairness principles. Through extensive experimentation and evaluation, we achieved promising results in accurately classifying images into their respective categories. Our model demonstrated robust performance in discerning subtle differences between real and AI-generated images, highlighting its potential utility in various applications, including image forensics, content moderation, and digital authentication. Continued Advancements in Model Robustness: Further refining CNN architectures and exploring advanced learning techniques to enhance classification accuracy across diverse datasets and evolving AI-generated image techniques. Real-Time Application Integration: Streamlining image classification models for real-time deployment in critical applications such as media verification, content moderation, and digital forensics to ensure rapid and reliable detection of fake images.

Ethical and Transparency Considerations: Developing frameworks that prioritize ethical considerations in the deployment of image classification technologies, including transparency in decision-making processes and adherence to privacy and fairness principles. Through extensive experimentation and evaluation, we achieved promising results in accurately classifying images into their respective categories. Our model demonstrated robust performance in discerning subtle differences between real and AI-generated images, highlighting its potential utility in various applications, including image forensics, content moderation, and digital authentication.

REFERENCE

[1] K. Roose, “An ai-generated picture won an art prize. artists aren’t happy,” The New York Times, vol.2, p. 2022, 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695, 2022.

[3] G. Pennycook and D. G. Rand, “The psychology of fake news,” Trends in cognitive sciences, vol. 25, no. 5, pp. 388–402, 2021.

[4] B. Singh and D. K. Sharma, “Predicting image credibility in fake news over social media using multi-

modal approach," Neural Computing and Applications, vol. 34, no. 24, pp. 21503–21517, 2022.
[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, "On the use of benford's law to detect gan-generated images," in 2020 25th international conference on pattern recognition (ICPR), pp. 5495–5502, IEEE, 2021.
[6] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10, IEEE, 2020.
[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, "Model inversion attack: analysis under gray-box scenario on deep learning based face recognition system," KSII Transactions on Internet and Information Systems (TIIS), vol. 15, no. 3, pp. 1100–1118, 2021.
[8] J. J. Bird, A. Naser, and A. Lotfi, "Writer-independent signature verification; evaluation of robotic and generative adversarial attacks," Information Sciences, vol. 633, pp. 170–181, 2023.
[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in International Conference on Machine Learning, pp. 8821–8831, PMLR, 2021.
[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, et al., "Photorealistic text-to-image diffusion models with deep language understanding," arXiv preprint arXiv:2205.11487, 2022.
[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, "Adapting pretrained vision-language foundational models to medical imaging domains," arXiv preprint arXiv:2210.04133, 2022.
[12] F. Schneider, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," arXiv preprint arXiv:2301.11757, 2023.
[13] F. Schneider, "Archisound: Audio generation with diffusion," Master's thesis, ETH Zurich, 2023.
[14] D. Yi, C. Guo, and T. Bai, "Exploring painting synthesis with diffusion models," in 2021 IEEE 1st International Conference on Digital Twins and Parallel Intelligence (DTPI), pp. 332–335, IEEE, 2021.