ISSN 2319-5991 www.ijerst.com
© 2026 IJERST. All Rights Reserved
Peer Reviewed, Referred & Indexed Journal

Research Paper

# IDENTIFY REAL VS AI GENERATED IMAGES USING CNN AND GRAD-CAM

**Mr.B.Kishore kumar**
*Dept of CSE(AIML)*
Vignana Bharathi Institute of Technology
Hyderabad, Telengana
botsakishor@gmail.com

**A.Gayathri**
Department of CSE (AI&ML)
Vignana Bharathi Institute of Technology
Hyderabad, Telengana
a.gayathrijanardhan02@gmail.com

**G.Vyshnavi**
Department of CSE (AI&ML)
Vignana Bharathi Institute of Technology
Hyderabad, Telengana
23p65a6605@vbithyd.ac.in

**D.Shivakumar**
Department of CSE (AI&ML)
Vignana Bharathi Institute of Technology
Hyderabad, Telengana
22p61a6643@vbithyd.ac.in

**Abstract:** In the recent developments, synthetic data has been developed and it has become easy to distinguish between the real and the created photograph. It is harder to navigate the ones created by a machine. The ease at which images created by AI can be found with the help of computer vision and other forms of informatics is significant as people must be able to trust and verify the information. Fake data have been created and compared with real pictures with the assistance of latent diffusion.

It applies computer vision algorithms, especially CNN2D, in sorting the images into groups. A more advanced CNN2D was developed. It is pretty good with the level of precision of 95.77. This will enable more photographs created by individuals to be discovered more often, as well as the growing problem of AI-generated content. These results indicate that computer vision is not a mere desirable attribute; rather it is a powerful tool that can identify genuine and counterfeit pictures that can be used to ensure that digital data is correct and valid.

**"KEYWORDS:** AI-generated images, Latent Diffusion Models, Generative AI,Fraud Prevention."

## 1. INTRODUCTION

AI has developed significantly in the past several years in terms of the art of creating faked photographs, transforming the manner in which human beings create and utilize photographs. The very first AI generated photographs had too many blemishes on their look, causing them to be instantly differentiated with the actual ones. Artificial models however gave way to more sophisticated models which over time could also depict images that were similar to actual ones so much that they could never be distinguished with images taken by human beings. The latest

improvement in this area is the LDMs, which could create realistic and high-quality images within a few seconds. The creative economy has always been revolutionized by the technical solution, as AI-generated art does not only compete with the art created by humans, but also has been awarded in art competitions [1].

Despite the significant developments in the field of image creation, there is also the emergence of the issue of trust, privacy, and authenticity due to the emergence of AI-mediated content. It is also negative to the current photo-checking technique as the LDM and other generative models can create realistic photos without referencing any real models. This has brought about many issues in areas like the journalism, digital media, and law enforcement where image quality is very significant in believing the content. The IT industry has continued to grow making people unable to differentiate between real and non-authentic photos. This can make the evidence in the images and news items less reliable [3, 4].

The ethical and social consequences of AI art are of concern to people because it appears to have fewer distinctions to the real world. Since current usage rates of technology can deliver quality images, there is a possibility that it could be abused. It could be misused by invasion of privacy, dissemination of fake information, and fraud. Deepfakes can be considered a serious concern, as they rely on artificial intelligence and are implemented to manipulate images and videos of real individuals with ill motives [5]. The fact that it is easier to commit fraud because of the possibility to create fake pictures and make them look like a real one makes the task of verifying the relevance of digital materials inefficient [6].

However, all has evolved since then, and we can now get to know what honesty and trustworthiness are in the digital world. The existence of the AI that can generate a photograph that seems to be a real one makes us wonder what a picture is and the role that technology plays in our life and the perception of the world. Since the frequency of using AI-generated photography is likely to increase, it becomes essential to establish a system that will help to explicitly display any deceptive picture to avoid fraud, fake news, and privacy breach [7]. Such are serious issues that need to be conquered to ensure that digital information becomes true and users remain convinced that picture information is true.

## 2. RELATED WORK

Image synthesis with the help of images and picture recognition with the help of AI is the subject of recent attention of scientists and scholars. This is attributed to the fact that simulated media is more realistic. Among the most significant issues that I would like to study is what is meant by prediction of synthetic images generated with various generation algorithms. Two of them are the text-to-image production and diffusion models.

DE-FAKE technique was developed by Sha et al. in 2022. It is able to identify and rank text to image models generating fake images. Their focus of interest is the way to figure out the difference between real pictures and the photos generated by AI, but not on how to determine the photos generated by AI. They consider the characteristics of created photos and join algorithms to find out who created an image which gives a useful tool of identifying fake material [8].

The article Corvi et al. (2022) examined the subject of so-called synthetic images created with the help of diffusion models and the need to realize what is

actually contained in these images. In addition, in this work, they also give us a means of identifying the fraudulent photos produced in the course of the diffusion process and showcase the originality of a synthetic image and a means of separating it with a real image. The essay shows the increasing importance of the creation of fake images in the fight of maintaining integrity of visual information in the digital media[9].

Li et al. (2020) were the authors of the paper that examined the strategies of differentiating images generated by deep networks through the analysis of color elements differences. Their mode of investigations of choice is founded on the color difference between original and counterfeit photographs. Surprisingly, this method is effective in the detection of false shots. The most trendy way of teaching with the help of AI systems is the process of changing the colors of data. It is typically applied to detect images generated by deep learning networks, to train and generate data in neural networks[10].

Their work by Krizhevsky and Hanto (2009) has led to the well-known AlexNet architecture that transformed the manner in which images are categorized. Even though they were not analyzing synthetic picture recognition, their work formed the basis of deep CNNs, which now play a pivotal role in the image recognition problem, regardless of whether the image is created by an AI. The application of AlexNet in categorizing images has seen further use in image detection studies, especially on synthetic media, in the future [11].

Gunning et al. (2019) developed a XAI methodology to reduce AI models and make them more comprehensible. The simplification of the AI systems mainly preoccupied them, which substantially affects the synthetic picture identification, especially in the aspect of expanding the ability of AI-based models to identify images produced by AI. XAI can be used to enhance the recognition algorithms [12] to understand more as to how the models are differentiating between real and fake photographs.

Ramesh et al. (2021) developed the Zero-Shot Text-to-Image Generation method, a method that converts text to images. Through this, models can make superior images of the language which is used to describe what is being discussed. This would go a long way in picture synthesis as one would be able to create a picture that looks real to the pen, which is the entirely text-based one. In their work, they discuss in great detail the fact that the number of generative art models is rapidly growing and that AI has the potential to produce images, which cannot be recognized by standard image verification tools. [13].

Diffusion of Text to Image Photorealistic In Saharia et al. (2022), the model is applied to combine a complex language processing system and an image diffusion system. It also involves the generation of language processing and image diffusion systems of complex languages. It is also likely that their style will lead to images that will be much more authentic than they would be in description. It generates pictures as well as mediating the difference between the known and the expressible. Generative models are already advancing towards the state of being able to produce images that are close to photos. The blurring of the AI and the AI-generated images is becoming more challenging. People have a very good understanding of language and therefore can be able to create images that can be very cohesive and applicable in the environment. This theorizes it becoming extremely hard to detect the fake media [14].

## 3. MATERIALS AND METHODS

The suggested approach takes the opportunity of the latest computer vision technologies to simplify the procedure of searching images created by artificial intelligence. The conversion of the CIFAR-10 dataset into a synthetic dataset is called latent diffusion. The artificial information is then used to produce a sequence of photographs which can be contrasted with actual photos. The photos are sorted by the CNN2D system [15]. However, the improved CNN2D that has already been refined to be more successful in identifying pictures is referred to as CNN2D2 [16]. The system also compares real images with the AI-generated ones and identifies them with the help of computer vision. The proposed system utilizes the following techniques in order to solve the increasing issue of identifying forgery photos and also to guarantee that the digital information is precise and complete.

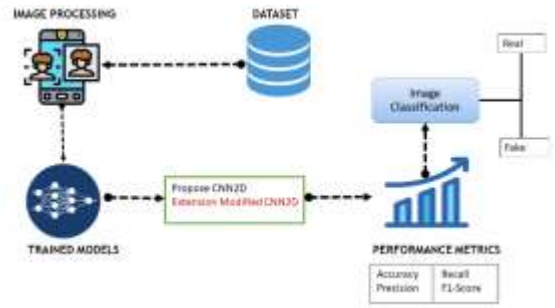



Fig.1 Proposed Architecture

The transmission of the images is followed by a picture (Fig. 1) that illustrates the processing of the genuine and the fake shots. This is among the designs of the system. The pictures are then stored in a database, which is later utilized in grouping the pictures. It is based on this information that two models are created, a CNN2D proposed CNN2D [15] and a custom CNN2D [16]. They are then defined to discriminate between authentic and fake images by a trained model paradigm. Model efficacy is determined based on accuracy, precision, memory and F1-score.

**i) Dataset Collection:**

The research will be based on fake and real information. Photographs that are difficult to capture, e.g. those with depth of field, shadow, and motion blur, do much better in comparison with SDM creation. Due to this fact, one would not be able to judge the difference between real and fake photos. The flaws and errors of the anatomy are present in some of the photographs, whereas the real images, like ripples in the water and shadows, are presented in other photographs.

**ii) Pre-Processing:**

The pretreatment measures in the study involve image editing software e.g. reducing, converting which is used to convert the image into grayscale. The last one is to picture the information in order to estimate the attractiveness of the picture. Training data is selected and randomly jumbled and pixel values are scaled to ensure as close as possible to the best performance model.

***Image processing:***

There has to be sufficient compilation of the information. The latter loads the images and balances them out in a way that they appear as identical. The other method that can be used to make sure that the data is balanced is a graph to demonstrate the distribution of the records in each category. After this, the photo pixel numbers are adjusted in such a way that they fall within the range of 0 to 1. Mixing up the data minimizes the chances of bias. This data is then separated into training and test sets and the model is developed and tested.

**iii) Algorithms:**

***CNN2D:*** CNN2D applies convolutional and pooling convolutions to obtain and arrange information in pictures. This application is credible because it has the ability to determine genuine and fake images depending on their marks [15].

***Modified CNN2D:*** The CNN2D+ CNN2D are more likely to feature pull and less prone to overfitting, as a result of the Global Average Pooling and Dropout layers. [16]. It is a simple installation that functions and aptly separates real and counterfeit photographs, which is even better.

## 4. RESULTS & DISCUSSION

**Accuracy:** The characteristics are stated in terms of the capability of the test to differentiate between healthy and sick individuals which is a sign of the quality of the test. To determine the accuracy of a test, the most suitable procedure is to establish the correlation between the number of true positives and the amount of true negatives in all cases that the test has been applied. mathematically it is:

$$"Accuracy = \frac{TP + TN}{TP + FP + TN + FN}(1)"$$

**Precision:** The measure of accuracy of diagnosing the appropriate cases is known as precision. Accuracy equation now appears as follows:

$$"Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}(2)"$$

**Recall:** To determine the capacity of a model to utilize all the notable examples on a particular class, the ML measure Recall is applied. This is a measure of the type of a model of a certain type of recognition event. In order to locate it you divide the number of correct positive observations by the correct positive forecasts.

$$"Recall = \frac{TP}{TP + FN}(3)"$$

**F1-Score:** The other strategy of measuring the correctness of an ML model is the F1 score. It is the effectiveness and the rate at which plans are executed. The accuracy of a model is the proportion of accurate predictions of a model on the total amount of data.

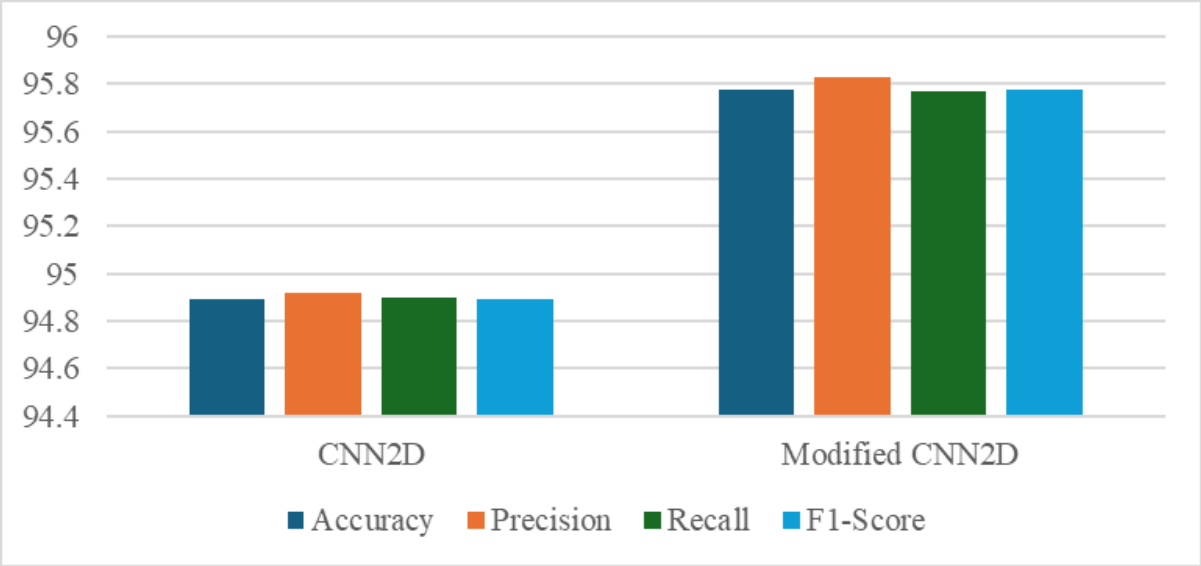
$$"F1\ Score = 2 * \frac{Recall\ X\ Precision}{Recall + Precision} * 100(1)"$$

***Table (1)*** presents the recall, F1-score, accuracy, and exactness of every method. It is also more precise than the old algorithms since it has been modified to make use of CNN2D. The tables provide the comparisons of the measures of the different algorithms in regard to their effectiveness.

"Table.1 Performance Evaluation Table"

| Algorithm Name | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| CNN2D | 94.890 | 94.921031 | 94.895695 | 94.889489 |
| **Modified CNN2D** | **95.775** | **95.829520** | **95.768225** | **95.773232** |

"Graph.1 Comparison Graphs"




The F1-Score, the accuracy, the precision, and the memory are illustrated in Graph (1) in the form of blue, orange, green, and sky blue lines. The Modified CNN2D Model had the highest results and it performed better than others. These data may be considered as graphs above.




Fig.2 Home Page

Figure 2 shows the user interface dashboard, comprising of a welcome message and the links to other websites.




Fig.3 Registration Page

By clicking the icons in the third picture on the top, you will be requested to provide your name, email address, phone number, and password.

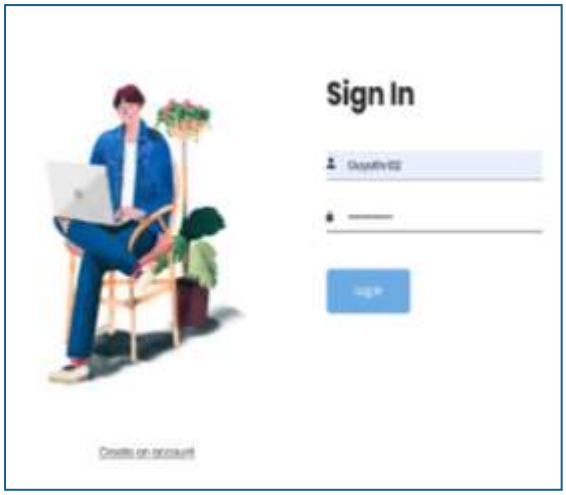



Fig.4 Login Page

In figure 4, there is a sign-in box containing a nickname field, a password field and a Create Account button.




Fig.5 Main Page

The navigation buttons are placed in the home page panel. Prediction, Graph, Notebook, and Signout are displayed in Figure 5.

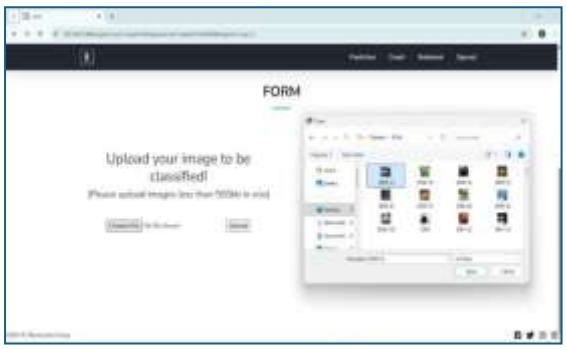



Fig.6 Upload Input Page

Figure 6 has an upload button and a coordinate input area at the left upper corner.




Fig.7 Predict Result for given input

The above outcome was derived by the use of the test data represented in Fig. 7.

## 5. CONCLUSION

Lastly, this work is an adequate attempt to explore the new threat of distinguishing between legitimate and fake images. This is increasingly finding it difficult because of the convergence of information. The system comprises a robust modified CNN2D which is useful in the clustering of photographs. The authors examine a counterfeit dataset constructed on the basis of CIFAR-10 dataset and latent diffusion to show inconsistencies between authentic photographs and some of the fakes produced by AI. Based on the accuracy of 95.77, it can be seen that the CNN2D algorithm performed significantly higher in the division of the two. The above performance model is among the means through which computer vision techniques can be implemented in confirming the genuineness of content and reducing fears of people regarding fake news. The results indicate that the CNN2D can be enhanced to enhance picture verification tools. It is the best way to search the AI-generated content and safely store the digital information.

In order to enhance this piece in the future, one can put more effort into the functionality of this model to guarantee that it can detect AI photos in a larger set of data, including high-resolution images and other ways of creating fake photos. They can also be improved by using more complicated techniques, like transfer learning and GANs. By applying the strategy in the real world, the fake news would be curtailed i.e. by the social media.

**REFERENCES**

[1] K. Roose, "An AI generated picture won an art prize. Artists aren't happy," New York Times, vol. 2, p. 2022, Sep. 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.

[3] G. Pennycook and D. G. Rand, "The psychology of fake news," Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.

[4] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using multi-modal approach," Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.

[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, "On the use of Benford's law to detect GAN-generated images," in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.

[6] D.Deb,J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10.

[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 Trans. Internet Inf. Syst., vol. 15, no. 3, pp. 1100–1118, Mar. 2021.

[8] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models," 2022, arXiv:2210.06998.

[9] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusio models," 2022, arXiv:2211.00680.

[10] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," Signal Process., vol. 174, Sep. 2020, Art. no. 107616.

[11] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[12] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang,"XAI—Explainable artificial intelligence," Sci. Robot., vol. 4, no. 37, Dec. 2019, Art. no. eaay7120.

[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen,and I. Sutskever, "Zero-shot text-to-image generation," in Proc. Int. Conf.Mach. Learn., 2021, pp. 8821–8831.

[14] 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, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic tex to-image diffusion models with deep language understanding," 2022, arXiv:2205.11487.

[15] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1205–1207.

[16] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A hybrid CNN LSTM model for video Deepfake detection by leveraging optical flow features," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2022, pp. 1–7.