[1]Dr. Rupak Sharma

[2]Dr. D Prabakar

[3]Aanchal Madaan

[4]Dr Devendra Kumar

[5]Dr.Makarand Upadhyaya

[6]Dr. Arvind Kumar Sharma

# Securing Social Media Imagery: GAN-Driven Encryption and CNN Analysis with DEA Protection

***Abstract: -*** Generative adversarial networks (GANs) transform low-dimensional random noise into a photorealistic picture. The use of such misleading pictures laden with irrelevant information on social media platforms might lead to serious and challenging issues. The goal of this study is to create and identify the affected GAN pictures. Additionally, enhanced GAN detection accuracy by preprocessing and segmentation feature extraction. This study examines the effectiveness of several learning-based methods for detecting image-to-image translation. Generative adversarial networks use deep learning approaches like convolutional neural networks for generative modelling and the DEA (Data encryption standard) is used to produce the encryption key. To accurately identify fake images, an effective picture forgery detector is necessary. Recent advancements in generative adversarial networks (GANs) have been focused on producing photorealistic pictures quickly and successfully. However, GANs can complicate visual forensics and model attribution. Data from one source and pictures from another can have a wide range of uses, including in fields like computer vision, video, and language processing. The investigation of some photos reveals that both conventional and deep learning detectors may reach up to 95% detection accuracy. However, only deep learning maintains good accuracy on compressed data. An article that explains how to spot GAN-generated false images on social media also provides context on GANs and the theoretical concepts underlying them.

***Keywords:*** GAN, DEA, CNN, Image-To-Image Translation, Deep Learning.

## I. INTRODUCTION

The advent of big data has made pictures vital information carriers. This has led in the realms of social media and the Internet of Things, to be concerned about the safety of digital photographs while they are being sent or stored. People and the media may be severely affected by security breaches using digital photos. The need to ensure the safety of digital photographs is important. Generative adversarial networks (GANs) and variational autoencoders are two examples of deep learning-based generative models that have recently found utility in the synthesis of photorealistic pictures and video. The synthetic face in pornographic videos, for example, may be created using the cycle GAN. In addition, GANs are capable of producing speech videos that have the fabricated facial expressions or content of any well-known politician, among other things, which might lead to major issues in society and politics. Consequently, a method for detecting false faces in images that is both effective and quick is required. This publication expands upon the earlier work to swiftly and successfully detect computer-generated false pictures. Given the extensive usage of deep neural networks in domains, it is also used for detecting GAN-generated false

[1]Department of Computer Applications, SRM Institute of Science and Technology, NCR Campus, Modinagar-201204, Ghaziabad, India. rupaks@srmist.edu.in

[2]Department of Computer Science and Engineering, Karpagam College of Engineering, Coimbatore, India. Affiliation with Anna University Chennai. Prabakaralam@gmail.com

[3]Department of Computer Applications, Global Group of Institutes, Amritsar, Punjab, India. aanchalmadaan20@yahoo.com

[4]Department of Computer Application,ABES Engineering College Ghaziabad Uttar Pradesh India. devendra.arya@gmail.com

[5]Department of Management & Marketing, University of Bahrain, College of Business Administration, Bahrain. makarandjaipur@gmail.com

[6]Post-Doc Researcher, Department of Creative Technologies & Product Design, National Taipei University of Business, Taipei City, TAIWAN. drarvindkumarsharma@gmail.com

pictures. This study tested a supervised learning method based on deep learning for detecting fraudulent images. Another aspect that has been overlooked is the fact that fraudulent picture detection is often seen as a simple case of a binary classification issue or model. As an example, the convolution neural network (CNN) was used to create the phoney picture detector in this scenario. Two models—generative and discriminative—make up Generative Adversarial Networks, the Model of Discrimination. Like a binary classifier, the discriminative model classifies photos into several categories. Image authenticity is verified by comparing it to a preset dataset. An Exemplary Framework for Producing Unsupervised generative modelling in machine learning involves identifying and acquiring patterns in incoming data to generate new instances that mirror the original dataset. Generative models may aggregate data distributions and create input-consistent variables.

This project aims to create and recognize GAN-made fake photos. GAN detection accuracy may also be improved by preprocessing and segmentation feature extraction. Test and compare multiple learning-based image-to-image translation detection methods in this experiment. One deep learning-based generative modelling approach is the generative adversarial network (GAN). Machine learning framework using generative and adversarial neural networks. Generative data structures include one network creating new data, whereas adversarial data structures require two networks.

A GAN's Generator and Discriminator networks function together. The generator's goal in this two-player game is to trick the discriminator into thinking it’s getting training set data. Deception is prevented by carefully distinguishing between authentic and false information. The two may acquire and train on complex data formats including audio, video, and image files together. The model learns to produce realistic pictures by first producing images with random noise. A picture is created by feeding a generator with input random noise that has been sampled using a uniform or normal distribution.

The discriminator learns to distinguish between genuine and false pictures by feeding it both the generator's output—fake images—and the training set's actual images. If the input is actual, then the output will be the probability of that. Several learning-based approaches for detecting image-to-image translation are evaluated in this study. If these assaults can be revealed, how much can be revealed, and under what terms?
To achieve this goal, examine several systems that have breakthroughs in image forensics and general purpose. Highly deep convolutional neural networks (CNNs) have been specifically trained for this purpose. The scenario when photos are shared on a social media platform, such as Twitter, will also be investigated. Since picture compression is standard practice when uploading images, it often hinders the effectiveness of forgery detectors, making this the most prevalent and difficult scenario to deal with.

The following sections will detail the methods used for detection, explain how the pictures used in the study were generated (concerning the technique described in), and the outcomes of the experiments.

## II. IMAGE-TO-IMAGE GAN TRANSLATION

A highly broad topic is image-to-image translation, which involves mapping pictures from one domain to corresponding images in another domain. Results for this problem have been remarkable when using GAN-based techniques. A huge number of matched pictures from both domains are typically used to train the network in most applications. Unfortunately, these correspondences and image pairings aren't always known in advance. Using GANs requires a new training strategy. Two nodes—a generator in the image-to-image network and a discriminator in the support network—make up the adversarial training paradigm. Both the discriminator and the generator are taught to detect and distinguish between actual and fake pictures. To compensate for the dearth of actual picture pairings, the suggested technique generates image pairs automatically that map onto each other. To get the output domain DB from the original domain DA, generator GA is used to convert input picture A, according to the design in Figure 1.

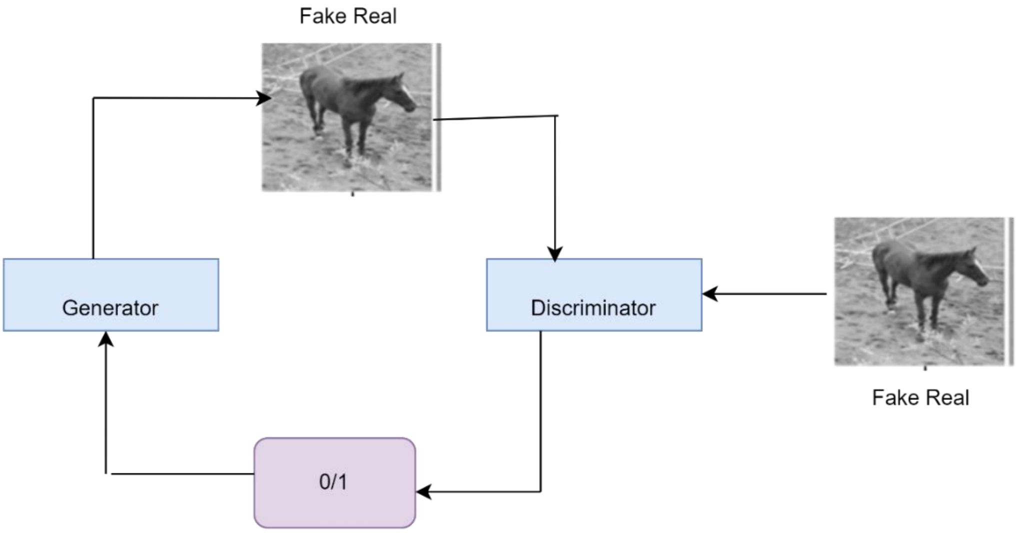

**Figure 1: cycle-GAN image-to-image translation**

After that, generator GB re-maps picture B to the initial domain DA. Two extra cycle consistency losses are used in this procedure to guarantee that A = GA(GB(A)), that is, that the two pictures constitute a matching pair, in addition to the typical adversarial losses linked with the discriminators, which guarantee that the produced images are well-suited to the new domain. The algorithm created the picture on the left, while the one on the right is a Google Maps download. No indications of potential manipulation are visible at first sight. Further instances, comparing the original pictures with their updated equivalents created by switching up the painting technique. A potent editing technique is changing the image's context, which allows for substantial alterations to be made without introducing discrepancies in lighting, shadows, or perspective. It may be rather difficult for those without specialized training to distinguish between these fakes and authentic photos. Therefore, it is evident that this method can generate trustworthy false news (Figure 2).

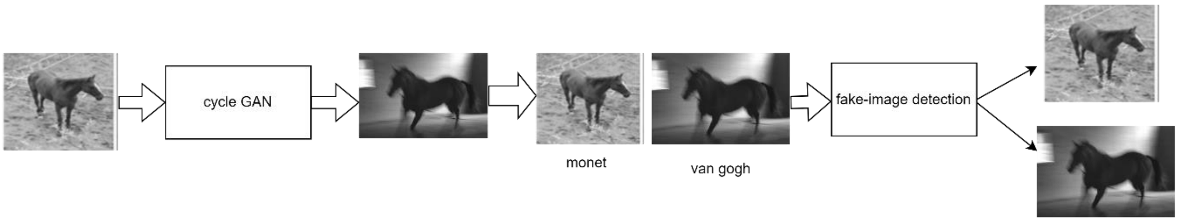

**Figure 2: Fake news on Twitter is distributed via a synthetic picture created by GAN-based image-to-image translation. Identifying false photos requires a trained classifier.**

## III. TESTED DETECTORS

Various approaches to identifying image-to-image translation are evaluated and contrasted. Some of these technologies were either developed for use in computer graphics picture detection or were suggested as a general approach to detecting image modifications. The rest rely on state-of-the-art convolutional neural network (CNN) designs that will tune to identify image-to-image translation. These models are highly deep and come from ImageNet. Furthermore, the same discriminator used by the authors in their GAN architecture is taken into consideration as a natural baseline. What follows is a synopsis of these approaches.

- **GAN discriminator:** The discriminator used its first thought as an element of the GAN. Referring to Figure 2, its structure has to be retrained using the dataset that is currently available to eliminate biases.

- **Steganalysis characteristics:** Although they were first suggested for use in steganalysis, they have since found usefulness in the detection of picture fraud, measuring the co-occurrences of detail micropatterns in high-pass residual pictures. In this case, a linear support vector machine (SVM) classifier is used after the best-performing model out of those that were suggested.

The DEA method is used to produce the encryption key. A key that may be used to encrypt the image generated by DEA after it evaluates the overall efficiency of the picture's pixels. In the second stage, the encrypted picture is put into a convolutional neural network (CNN) for advanced encryption. Preserving secrecy, trustworthiness, and authenticity is the primary goal of image security. An encryption method is one of several accessible approaches and ways of creating safe pictures. By default, encryption is a procedure that uses a key to convert a picture into a cryptic image. Another option is to apply a decryption technique to the cypher picture; this is often just the opposite of the encryption process, but it allows the user to get the original image. Figure 3 shows the primary image figure 4 depicts the result of an encrypted process, which is an encrypted picture. The process begins with a main image, which the user then uses to create a secret image using an encryption approach. Or, when the recipient receives this encrypted picture, he uses the decoding handle and restores the initial data. Figure 5 shows the picture that was recovered.

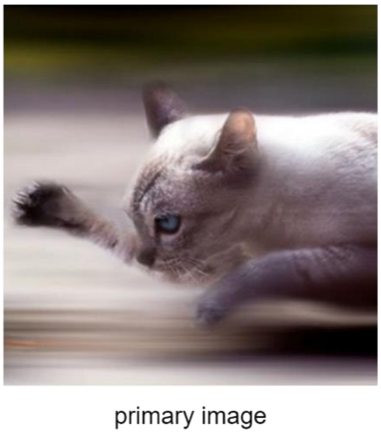

**Figure 3: Primary Image**

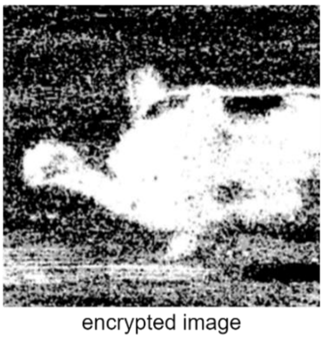

**Figure 4: Encrypted Image**

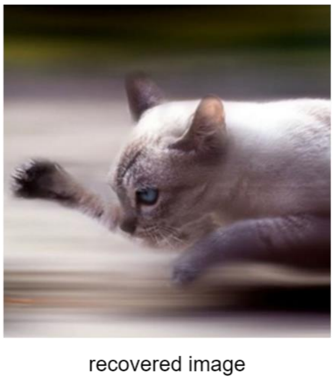

**Figure 5: Recovered Image**

The following are examples of some of the suggested approaches to identifying GAN architectures (figure 6) are the basis of computer graphic pictures and others have utilized the dataset to train adversarial networks.

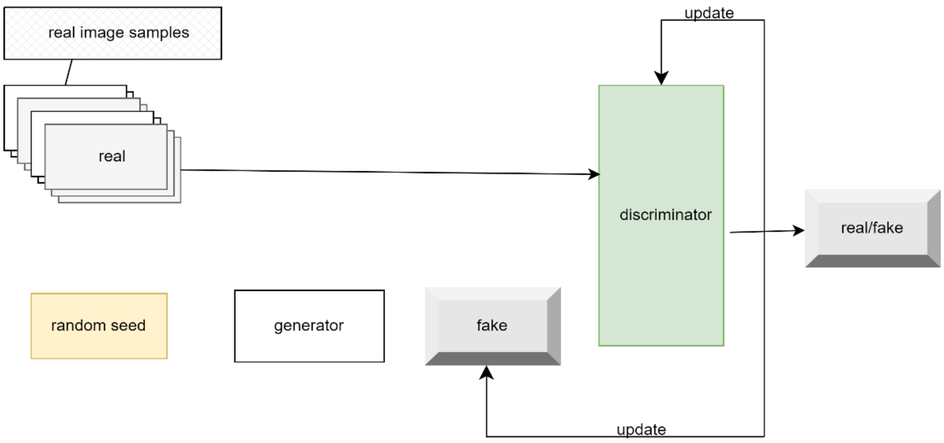

**Figure 6: GAN Architectures**

The dataset includes both genuine and counterfeit images for each category. For instance, the photos from social media are used to train GAN, and the network generates corresponding fake images. To begin, use the dataset to train adversarial networks. There are both authentic and fabricated pictures in the dataset for every category. There are primarily two steps to the training process. Training the discriminator and then "freezing" the generator (training as fake) is the first step. The network is designed to merely forward packets.

The next step is to train the generator while simultaneously freezing the discriminator, before a GAN is trained, and identify the issue. The goal here is to create visuals that do not exist. Convolutional neural networks (CNNs) will serve as both the generator and discriminator in a GAN, which means that its design must be defined. Once the original photos, the discriminator recognises them as genuine.

Use the generator to create bogus photos and allow the discriminator to accurately identify them as phoney. Instruct the generator to use the discriminator's output. The discriminator's predictions are used to train the generator. For more precise results, repeat these procedures many times. The basic idea behind GANs is a minimax game, in which the loss-minimizing discriminator and the loss-maximizing generator compete for data that stands out. The mathematical expression of GAN's operation is as follows:

The function V= (D, G) =Ex ~ p data(x) [log D(x)] is defined as the equation. Ez raised to the power of p data (z) [log (1-D (G (Z)))] < in which

"Discriminator" and "G-Generator" Extraction of X-sample from actual data and Z-sample from a generator

D(X)-Discrimination network, Pdata(x)-Generator distribution, and pdata(z)-Real data distribution Internet of G(Z)-Generators. By computing co-occurrence matrices on an image's RGB channels, discover GAN images. By applying various filters to the picture and calculating the difference, these matrices are often generated as image residuals.

In this study, however, bypass this problem by computing co-occurrence matrices on the picture pixels on the red, green, and blue channels independently. These matrices are fed into a convolutional neural network, which uses them to acquire useful features.

1. Step one involves obtaining a 3x256x256 tensor by computing co-occurrence matrices on RGB channels. Afterwards, it undergoes processing by a multi-layer deep convolutional neural network, which consists of the following layers: convo, ReLU, max pooling, dense, and sigmoid. Picture data need to be stored in the input layer of CNNs. There is a three-dimensional matrix that represents the image information. Because it is inside this layer that picture characteristics are retrieved. The convolutional layer is also known as the feature extraction layer. To turn all negative values into zero, the Convo layer includes ReLu activation. The rectifier unit, or ReLu, is the most often used operation on the outputs of convolutional neural networks (CNN) neurons (figure 7).

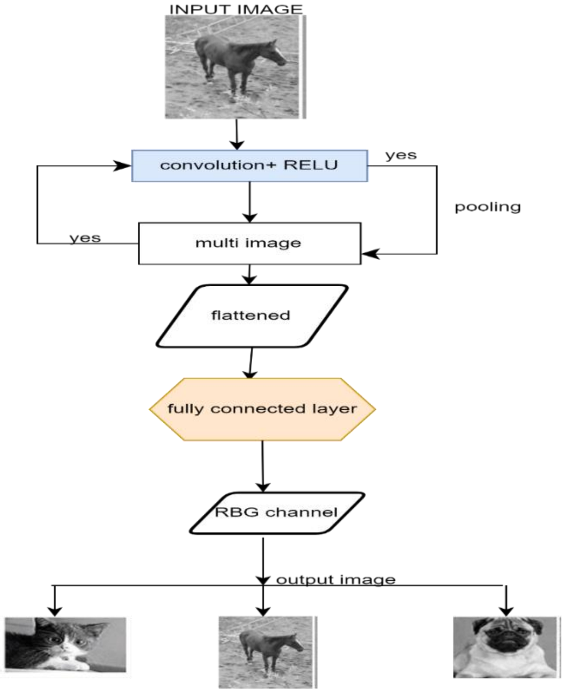

**Figure 7: Convolution neuronal network**

2. Following the convolution layer, a pooling layer is used to decrease the spatial volume of the input picture. The output of a sigmoid layer is constrained to the interval (0,1) since the input is transformed using a sigmoid function. Neurons, weights, and biases make up a fully linked layer. Its primary function is to divide the photos into many categories by training. A single hot-encoded label is included in the output layer.

## IV. EXPERIMENTAL RESULTS

### 4.1 Data set

This experiment made use of the Cycle GAN dataset, which contains a formatted collection of naturally occurring pictures. A minor number of photos were chosen at random from this dataset. From horses to zebras, from summer to winter, and from pictures to paintings, among other things, this dataset comprises unpaired image-to-image translations produced utilizing a cycle-consistent GAN framework. Picture to image, collect samples of various regions to create a massive dataset translation. There are both authentic and fabricated pictures in the collection. It can be seen in action in the "horse to zebra" subset, which contains both the real photos of horses and zebras that were used to train the GAN and the matching fakes that the GAN created after training. The first set of topics covered the big collection of natural picture translations, which includes horse to zebra. The second collection is concerned with creating pictures using labelled cityscape maps; in this instance, the dataset only contains actual, one-time-generated photographs.

A collection of records constructed a massive dataset of samples from several categories of image-to-image translation using publicly accessible code to train and test the detectors being compared. There are both authentic and fabricated photographs in the collection for every category. For instance, in the apple and orange subset, both the authentic apple and orange photos were used to train the GAN, as well as the matching fakes that the GAN created after training. All categories are shown in figure 8.

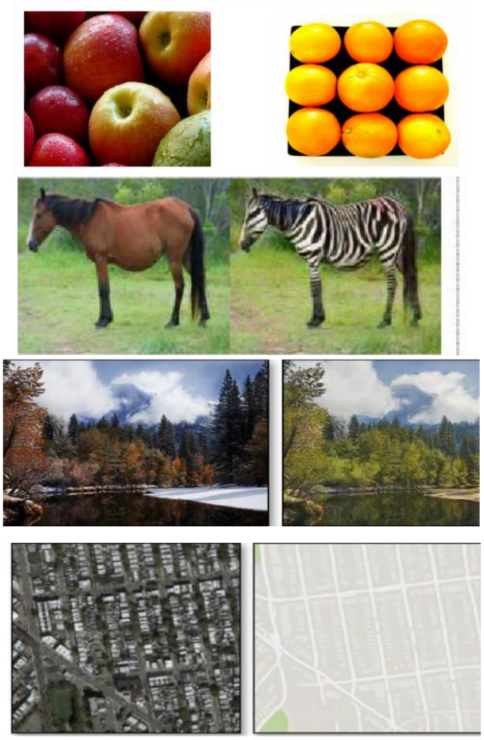

**Figure 8: The collection includes image-to-image translation categories. Parentheses indicate the amount of authentic and fraudulent photos accessible.**

Based on their contents, these alterations may be further categorized. Included in the first category are apple2image, horse2zebra, and winter2summer, all of which deal with the translation of real-life photographs. Second, there's the category of pictures created from tagged satellite photos, building facades, and cityscape maps. Here, the dataset just contains produced and actual photographs; neither the real nor the created handcrafted label maps are taken into account. Finally, there's the category that includes making paintings out of photographs and vice versa. In all, there are almost 36,000 256×256 colour photos in the collection.

### 4.2 Protocol

LOMO (leave-one-manipulation-out) is a technique that is used to evaluate performance in real-world scenarios when the manipulation is unknown a priori. Consequently, throughout each cycle, reserve all photos belonging to a certain category for validation and utilize the rest for training. To generalize across manipulations, the classifier should learn patterns shared by all pictures created by this technique, rather than adapting to individual translation aspects.

### 4.3 The Scenario 1

Table 1 includes the results of each manipulation type and their average accuracy, weighted by the number of samples, in the final column for all procedures that were investigated. The average accuracy is the greatest among the methods that were studied.

Except for winter and summer, this shallow network achieves almost flawless categorization on average. Still, Xception Net and the hand-crafted features are two of the deep networks that provide excellent outcomes. Additionally, publish the average categorization accuracy for every alteration in the last row. Specifically, winter2summer and map2sat, to identify the most extreme instances. To mimic snow or vegetation in real settings without significantly altering the picture structures, the former uses an image-to-image translation that mostly includes lighting alterations and colour swaps. The second one, meanwhile, uses a named map as input and creates a satellite picture from the ground up that looks a lot like Google Maps. There are almost no visible artefacts in these photos, and they seem quite genuine. On the other side, horse2zebra and a few style transfers in the painting are the most noticeable modifications. Visual artefacts are created with new picture structures in all three situations; in the first, they are quite obvious, and in the paintings, they are less obvious. Obscured to some extent by the appealing aesthetic. Here, a human observer would have a hard time seeing the false photos, while an automated classifier would have no trouble at all.

**Table 1 Original Uncompressed Data- Results**

| Accuracy | Ap2or | Ho2zeb | Win2sum | Map2sat | Average |
|---|---|---|---|---|---|
| **Steganalysis feat.** | 99.94 | 99.45 | 67.24 | 89.10 | 88.93 |
| **GAN discr.** | 70.85 | 91.78 | 53.32 | 91.45 | 76.85 |
| **DenseNet** | 80.06 | 96.78 | 68.69 | 79.31 | 81.21 |
| **Inception Net v3** | 85.96 | 95.79 | 59.77 | 71.55 | 78.26 |
| **Xception Net** | 96.92 | 99.17 | 77.75 | 77.80 | 87.91 |
| **Average** | 86.74 | 96.59 | 65.35 | 81.84 | 82.63 |

### 4.4 The Scenario 2

Although the previous experiment yielded promising findings, they should be interpreted with care. It seems very improbable that the original malicious user-generated picture is immediately accessible. The most probable place for its distribution is on a social media platform, where it might potentially get widespread attention and eventually go viral. Automatic compression happens to most submitted photos, erasing the delicate patterns that classifiers rely on. Thus, when it comes to picture forgery detection, resilience is a big matter. Table 2 displays the outcomes when compression is present for the same detectors that were tested before that is, the classifiers are still trained on original samples but tested on compressed ones. The compression method used is Twitter-like, meaning it mimics the JPEG compression that is used when an image is tweeted. This includes the quantization table, chrominance sub-sampling, and quality factor. Most detectors severely underperform when compared to this basic routine operation, particularly those that rely on characteristics that are either handmade or have poor computer systems that use neural connections. With an accuracy of 87.17%—just 7% lower than the uncompressed case—CNN demonstrates superior robustness, particularly XceptionNet. This suggests that these networks depend on qualities other than textural micropatterns that can withstand compression.

**Table 2 Training Mismatches Twitter-like Compressed Images**

| Accuracy | Ap2or | Ho2zeb | Win2sum | Map2sat | Average |
|---|---|---|---|---|---|
| **Steganalysis feat.** | 51.21 | 51.13 | 51.01 | 51.01 | 51.09 |
| **GAN discr.** | 51.16 | 56.25 | 50.92 | 51.28 | 52.4 |
| **DenseNet** | 77.60 | 91.33 | 55.18 | 62.70 | 71.70 |
| **Inception Net v3** | 84.86 | 87.93 | 53.68 | 61.01 | 71.87 |
| **Xception Net** | 91.88 | 97.10 | 53.58 | 52.35 | 73.73 |
| **Average** | 71.34 | 76.74 | 52.87 | 55.67 | 64.08 |

### 4.5 The Scenario 3

The previous experiment performed, of course, accounts for the worst-case scenario, when the training and test sets were not a good fit. Training the classifiers directly on compressed photos makes sense when attempting to identify social media fakes. In this situation, the findings are detailed in Table 3. Because some information is crucial for detecting the fakes that were lost during the lossy compression and cannot be retrieved, the steganalysis features now function quite well, but they are still over 10% worse than the uncompressed scenario. As always, XceptionNet delivers the best results, this time with an accuracy of 89.03%.

**Table 3 Twitter-Compressed Image Results**

| Accuracy | Ap2or | Ho2zeb | Win2sum | Map2sat | Average |
|---|---|---|---|---|---|
| **Steganalysis feat.** | 80.40 | 91.03 | 57.67 | 70.40 | 74.87 |
| **GAN discr.** | 64.30 | 92.09 | 52.91 | 80.35 | 72.41 |
| **DenseNet** | 79.28 | 94.45 | 67.95 | 81.46 | 80.78 |
| **Inception Net v3** | 79.61 | 96.24 | 65.55 | 64.85 | 76.56 |
| **Xception Net** | 94.53 | 94.79 | 68.08 | 68.98 | 81.59 |
| **Average** | 79.62 | 93.72 | 62.43 | 73.20 | 77.24 |

## V. CONCLUSION

Research on picture manipulation detection using GAN-based image-to-image has been reported. Although several detectors exhibit strong performance on original photos, a few exhibit significant limitations when presented with compressed images such as those used by Twitter. When there is a training-test mismatch, deep networks, and Xception Net, in particular, maintain their robustness and continue to function relatively well. In addition to expanding the analysis to include additional manipulations and detectors, future research will investigate the effectiveness of different methods of other synthetic picture generators, following transfer learning. Research in the future will focus on studying the cross-method performance of other synthetic picture generators, maybe after the transfer learning, in addition to expanding the analysis to include additional manipulations and detectors.

Additionally, evaluate performance in real-world circumstances using various social networks. Ultimately, GAN is an excellent technology in society as well as medicine. This helps to identify bogus data from actual data. The original photos must be analyzed rapidly. To identify GAN-generated false pictures, a new approach is presented in this research that uses pixel co-occurrence matrices and advanced analytics. To differentiate between actual and GAN-generated false pictures, co-occurrence matrices are calculated on the image's colour channels and then trained with a deep convolutional neural network (CNN). An analysis of a few images shows that conventional and deep learning detectors are equally capable of achieving 95% detection accuracy. This method is successful and may be applied to other GAN datasets, as shown by experiments. In the future, explore methods to accurately identify the modified pixels in GAN-generated counterfeit images.

## REFERENCE

[1] Rajasoundaran, S., et al. "Secure and optimized intrusion detection scheme using LSTM-MAC principles for underwater wireless sensor networks." *Wireless Networks* (2023): 1-23.

[2] Li, Mengfang, et al. "Medical image analysis using deep learning algorithms." *Frontiers in Public Health* 11 (2023): 1273253.

[3] Lv, Zhihan, et al. "Behavioral modelling and prediction in social perception and computing: A survey." *IEEE Transactions on Computational Social Systems* (2022).

[4] Zhao, Ying, and Jinjun Chen. "A survey on differential privacy for unstructured data content." *ACM Computing Surveys (CSUR)* 54.10s (2022): 1-28.

[5] DAIRI, ABDELKADER, and YING SUN. "A Semi-Supervised Modulation Identification in MIMO Systems: A Deep Learning Strategy."

[6] Qu, Youyang, et al. "Privacy Preservation in IoT: Machine Learning Approaches." (2022).

[7] Prasad, Rajesh S., et al. "FCM with Spatial Constraint Multi-Kernel Distance-Based Segmentation and Optimized Deep Learning for Flood Detection." *International Journal of Image and Graphics* (2023): 2450041.

[8] Liu, Wenkang, et al. "BFG: privacy protection framework for internet of medical things based on blockchain and federated learning." *Connection Science* 35.1 (2023): 2199951.

[9] Rashid, Md Tahmid, Na Wei, and Dong Wang. "A survey on social-physical sensing." *arXiv preprint arXiv: 2104.01360* (2021).

[10] Qu, Youyang, et al. "Privacy Preservation in IoT: Machine Learning Approaches: A Comprehensive Survey and Use Cases." (2022).

[11] Nicolazzo, Serena, et al. "Privacy-Preserving in Blockchain-based Federated Learning Systems." *arXiv preprint arXiv: 2401.03552* (2024).

[12] Gao, Longxiang, et al. *Privacy-preserving in edge computing*. Springer, 2021.

[13] Sood, Pallavi, et al. "Review the role of artificial intelligence in detecting and preventing financial fraud using natural language processing." *International Journal of System Assurance Engineering and Management* 14.6 (2023): 2120-2135.

[14] Liu, Xiaolan, et al. "Distributed intelligence in wireless networks." *IEEE Open Journal of the Communications Society* (2023).

[15] Chen, Mingzhe, et al. "Distributed learning in wireless networks: Recent progress and future challenges." *IEEE Journal on Selected Areas in Communications* 39.12 (2021): 3579-3605.

[16] Mewada, Shivlal, et al. "Smart diagnostic expert system for defect in forging process by using machine learning process." Journal of Nanomaterials 2022 (2022).