DOI: 10.24996/ijs.2025.67.1.40

ISSN: 0067-2904

Digital Forensics Method for Fake Images Detection Using GAN Algorithm Based on Watermarking Technique and Image Content

Shahlaa Mashhadani[1], Wisal Hashim Abdulsalam[1*], Wisam Abed Shukur[1], Mohammed S. H. Al-Tamimi[2]

[1] *Computer Department, College of Education for Pure Science/Ibn-Al-Haitham, University of Baghdad, Baghdad, Iraq*

[2]*Computer Science Department - College of Science - University of Baghdad - Baghdad – Iraq*

Received: 8/10/2024 Accepted: 10/1/2025 Published: xx

Abstract

The proliferation of manipulated multimedia content poses a significant threat in an era heavily reliant on social networks as primary information sources. Despite numerous countermeasures targeting specific attack types, the seamless nature of image manipulation challenges the differentiation between authentic and altered visuals. This study aims to detect fake images generated by StyleGAN2-ADA using watermark analysis and image content analysis techniques. The first experiment evaluates the performance of watermarking techniques in the spatial (Least Significant Bit (LSB)) and frequency (Discrete Cosine Transform (DCT)) domains using real-life imagery. Then, watermarked images are used as input to the StyleGAN2-ADA model to generate synthetic counterfeits. The second experiment assesses the effectiveness of content-based analysis techniques in distinguishing between authentic and forged images, including Error Level Analysis (ELA), perceptual hashing, and a pre-trained Convolutional Neural Network (CNN) model. Finally, the third experiment integrates the findings from the two previous experiments to provide a reliable determination of image authenticity. The results show that by leveraging watermark-based and content-based detection, the proposed framework achieves high accuracy in identifying fake images generated by StyleGAN2-ADA.

Keywords: Digital Forensics, Fake Images Detection, StyleGAN2-ADA, Watermarking Technique, Image Content Analysis

طريقة التحليل الجنائي الرقمي للكشف عن الصور المزيفة باستخدام GAN القائمة على تقنية العلامات المائية ومحتوى الصورة خوارزمية

شهلاء المشهداني[1], وصال هاشم عبد السلام[2*] ، وسام عبد شكر[3] , محمد التميمي[4]

[1,*3،2] قسم الحاسبات، **كلية التربية للعلوم الصرفة/ ابن** الهيثم، **جامعة**

بغداد، بغداد، العراق

[4]قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

[*]Email: wisal.h@ihcoedu.uobaghdad.edu.iq

الخلاصة

يُشكل انتشار المحتوى الوسائطي المتعدد الذي تم التلاعب به تهديدًا كبيرًا في عصر يعتمد بشكل كبير على شبكات التواصل الاجتماعي كمصدر رئيسي للمعلومات .وعلى الرغم من وجود العديد من الإجراءات المضادة التي تستهدف أنواعًا معينة من الهجمات، إلا أن طبيعة التلاعب بالصور بسلاسة تشكل تحديًا في التفريق بين المرئيات الأصلية والمعدلة .تهدف هذه الدراسة للكشف عن الصور المزيفة التي تم إنشاؤها باستخدام StyleGAN2-ADAباستخدام تقنيات تحليل العلامات المائية ومحتوى الصورة .تقيم التجربة الأولى أداء طرق العلامات المائية في المجالين المكاني (أقل بت أهمية(LSB)) والترددي (تحويل جيب التمام المنفصل(DCT)) باستخدام صور واقعية .ثم يتم استخدام الصور المائية كمدخلات لنموذج StyleGAN2-ADA لإنشاء تزييفات صناعية .تقيم التجربة الثانية فعالية تقنيات تحليل المحتوى، بما في ذلك تحليل مستوى الخطأ، والهاش ونماذج الشبكة العصبية التلافيفية (CNN) المدربة مسبقًا، في التمييز بين الصور الأصلية والمزورة .أخيرًا، تدمج التجربة الثالثة نتائج التجربتين السابقتين لتوفير تحديد موثوق لمصداقية الصورة .تظهر النتائج أن الإطار المقترح، الذي يستفيد من الاكتشاف القائم على العلامات المائية والمحتوى، يحقق دقة عالية في تحديد الصور المزيفة التي تم إنشاؤها بواسطةStyleGAN2-ADA .

1. Introduction

As per the International Telecommunication Union (ITU), by the end of 2019, approximately 4.1 billion people worldwide have been using various online tools, with us experiencing widespread use of Internet services and a significant rise in social media platforms growth - all happening during these digital times. With genuine content there is a troubling increase in intentional deception perpetuating misinformation. Editing software and AI have improved greatly, meaning that all visual fakes now look almost real. It is this trend that carries a significant threat when the manipulated content becomes digital evidence in law and forensic investigations [1].

At the same time, the discipline of digital forensics has gained importance in verifying these altered or fake images. Various state-of-the-art technologies, such as Generative Adversarial Networks (GANs), watermarking methods, and image content analysis, are combined to improve the efficacy of existing methodologies to unveil manipulated images. More and more organizations, from news agencies to legal firms to intelligence organizations, need to authenticate multimedia products. In multimedia forensics, researchers have proposed methods to analyze images, searching for evidence of tampering, fingerprinting (direct tracing of manipulation traces), and inconsistencies in imaging cues (e.g., lens aberrations, sensor noise). Deep learning-based methods have recently become capable of attaining impressive results in revealing fake parts in images [2-4].

However, with deepfakes, the spreading of manipulated multimedia has become a real threat in these heavily social network-reliant periods because of reliance on it as the main source of information. Also, in image and video manipulation, it deals with specific actions performed on digital content using editing software tools like Adobe Photoshop, GIMP, PIXLR, or even AI. Such methods include "copy-move," where a region of an image is copied and then pasted within the image. With advances in editing tools, fake images have become qualitatively much better and, to the naked eye, cannot be differentiated from the original. Additional post-processing manipulations, such as JPEG compression, changes in brightness, or equalization, may further reduce traces of manipulation and make the process of detection harder. Because there have been too many proposed countermeasures that emphasize a single type of attack, the nature of the manipulation of an image remains challenging to distinguish an authentic one from a manipulated one. Instead, integrating GANs with watermarking techniques and image content analysis will provide a holistic solution to such issues. GANs provide an opportunity to generate content, while watermarked information remains traceable due to its nature.

Therefore, each layer allows for the detection and reduction of manipulated visual propagation. It has been a landmark in digital forensics and a structural and subtle approach to the authentication and verification of images in the rising tide of manipulated images [5-8].

The present study proposes an intelligent approach in digital forensics for identifying fake images generated through Generative Adversarial Networks via the StyleGAN2 algorithm, embedded with hidden text watermark approaches and image content analysis. Deep learning models have completely revolutionized the trend in creating and editing images; among them, the emergence of GANs is one such evolution. Hence, this paper utilizes ADA, an improved version of StyleGAN2. This technique harnesses the capabilities of GANs for generating realistic fake images by embedding hidden text watermarks in images imperceptibly to serve as a unique identifier for authenticity verification. Further, it aids forensic experts in tracing image origins and validating authenticity through hidden watermarks. The detection process is improved by analyzing image content for inconsistencies using various metrics. The integration of watermarking and content analysis enhances fake image detection.

The proposed digital forensics pipeline will help mitigate a wide range of challenges arising from forged images across various settings, from fighting misinformation and retaining integrity in photojournalism to examining digital evidence in legal settings. This work is one contribution to the fast-growing study area of digital forensics and cybersecurity. It reflects an attempt to secure the credibility of visual information in an age when it is becoming increasingly difficult to discern what is real and what is deception.

This paper is structured as follows for the remaining sections: Section 2 discusses the theoretical part. In Section 3, the methodology of the proposed algorithm is presented. Section 4 is for results and discussion. Finally, in Section 5 our findings are concluded with areas for future development at the end of the paper.

2. Forensics and Fake Image Detection

Forensics and fake image detection are the prime elements in establishing the credibility of a digital image. In this regard, forensics involves using scientific techniques and instruments to examine the image to authenticate it. It includes metadata analysis techniques, detection algorithmic techniques, techniques used in digital image analysis, advanced technologies AI and DL, and verification of the source [5, 9].

Metadata examination involves the analysis of embedded information within an image file. It comprises the examination of timestamps, location information, and editing history for any inconsistencies that might point to manipulation. Algorithmic Detection: Detection by algorithms uses a wide range of statistical and ML models to flag images based on various irregularities related to editing software.

Digital Image Analysis involves checking the pixels, color gradients, and patterns of an image for any irregularities that might show evidence of tampering, cloned,regions or splicing of several images.

Source Verification helps trace the roots of an image to validate its authenticity by cross-checking the information against other reliable sources or by using reverse image searches. Such techniques, in turn, can be used to verify digital images to retain the accuracy and trustworthiness of information disseminated through digital networks. These approaches are basic in limiting sham images and, consequently, in protecting the authenticity of digital content while reducing the harmful effects of misinformation in discourse and decision-making processes [10]. The two main categories of image forensics detection are classical and modern techniques.

2.1 Classical detection techniques

There are several traditional techniques and methods for detecting manipulations, alterations, or inconsistencies in digital images. Among the front-runner methods of traditional image forensics are the analysis of the metadata of an image file. This would provide information concerning timestamps, edit history, and camera settings. The presence of any irregularities or lack of coherence in such information could be indicative of some manipulation or tampering. Other techniques also involve the analysis of the image itself, such as ELA, noise analysis, and analyzing the inconsistencies in lighting or shadows. The ELA would check the level of compression of different areas in the image because different editing actions will exhibit different types of compression artifacts [11].

Another classical approach to finding the hidden information or detecting tampering of an image is through steganalysis. This technique will look for hidden data in the image or information that might have been embedded, removed, or altered. In addition,, an image's statistical properties can also reveal inconsistencies. For example, statistical features such as the correlation between neighboring pixels and color histogram variations may identify possible tampering. Traditional methods of image forensics are relatively well-rooted, but they tend to depend highly on manual observation and can have limited functionality in terms of complex tampering or deepfakes. They are nevertheless key tools within the greater context of image forensics, which are the complementary advanced AI techniques to enhance the authentication of digital imagery and the need to fight the spread of fake images [12-15].

2.2 Modern detection techniques

The learning-based techniques involved using ML, DL, and AI to detect and classify fake and manipulated images. Unlike classical forensic techniques that perform rule-based analyses to detect image tampering, learning-based detection relies on algorithms trained on large datasets to automatically identify patterns, anomalies, or inconsistencies in an image that may signal image tampering. The training of ML algorithms -essentially deep neural networks -fed varied datasets of authentic and manipulated images has thus enabled them to extract features and patterns from the data while differentiating between the real and manipulated images [16, 17].

Convolutional Neural Networks were widely used in learning-based detection; these networks analyzed image features through a number of layers for changes in pixel value, texture, or abnormal patterns that are different from the norm. Transfer learning, where previously trained models are adapted for a different task, was also used to enhance the efficiency of the detection algorithms [18, 19]. On the other hand, Generative Adversarial Networks (GANs) were applied in learning-based detection, comprising two neural networks, the generator and the discriminator, that worked against each other. Meanwhile, the generator attempted to produce fake images, and the discriminator aimed to distinguish between real and fake images. This adversarial process resulted in improved detection capabilities as both networks continuously improved [20].

Learning-based detection provides several advantages compared to classical techniques. Firstly, the method can learn evolving manipulation techniques and cope with more involved alterations, such as deepfake videos or highly sophisticated image forgeries. It depends on the quality and diversity of the training data and on the continuous evolution of the detection model, such that new, sophisticated image manipulation techniques can be detected [21].

3. Material and Methods

The proposed work's block diagram appears in Figure 1 for fake image detection.

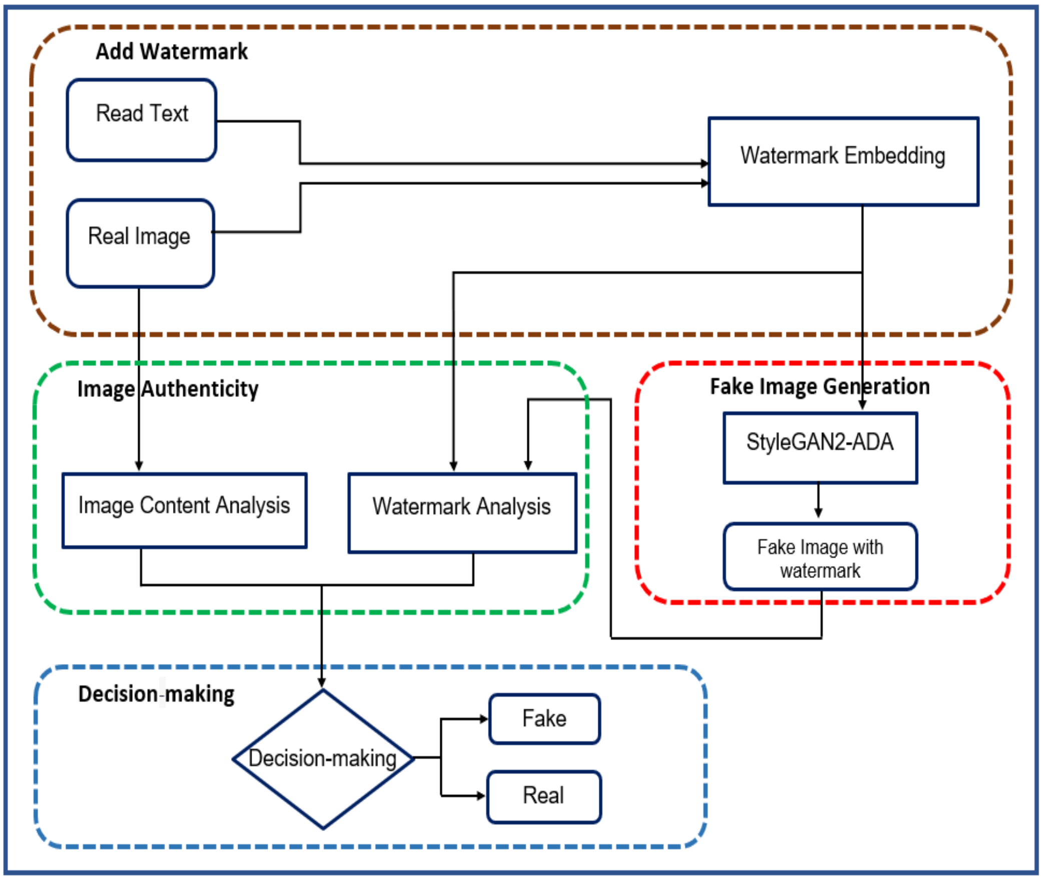


Figure 1: Block diagram of the proposed framework

As shown in the above figure, three experiments were applied, which are:

Experiment 1: The use of watermarking methods in different domains using real-life imagery.
Experiment 2: The use of different techniques for content image analysis.
Experiment 3: Applying a fusion approach of the above experiments (1 and 2) to improve the performance.

Experiment 1 Methodology

Unfortunately, adding a watermark to an image to prevent or aid in fake image detection typically requires some level of technical intervention or software manipulation. Once the watermark is added, it can serve as a visual indicator of authenticity. This won't necessarily prevent tampering or manipulation, but it can act as a signal to verify the originality of the image. In this work, hidden watermarking over images is chosen in the context of fake image detection due to several factors such as robustness, imperceptibility, embedding capacity, verification ease, and forensic interpretability. However, watermarks are used in two domains, the spatial domain and the frequency domain.

The purpose of this experiment is to assess the performance of the watermarking methods across different domains to understand their effectiveness in detecting tampered images generated by StyleGAN2-ADA. Two different techniques in different domains, which are Least Significant Bit (LSB) and Discrete Cosine Transform (DCT), are employed to embed the text in real images, as illustrated [22]:

Algorithm 1 LSB in Spatial Domain

```
Input: Text, Image
Output: Watermarked image
Select an image with any text.
1.    Convert the image to PNG format.
2.    Convert the image to grayscale if it's not already in grayscale.
3.    Crop the image to focus on the face, then resize it to 256*256.
4.    Read the watermark text (5000 characters).
5.    Convert the watermark text to bits.
6.    Embed the watermark bits into the image.
7.    Save the watermarked image.
```

Regarding the frequency domain, DCT is used due to its robustness to attacks, which makes it suitable for embedding and detecting watermarks in images.

Algorithm 2: DCT in Frequency Domain

```
Input: Text, Image
Output: Watermarked image
1.    Read an image with any text (5000 characters).
2.    Convert the Image to YCrCb Color Space
3.    Slice the image into blocks (each block size was 8*8)
4.    Apply the DCT to each block.
5.    Read the watermark text (5000 characters).
6.    Embed Watermark Bits in DCT Coefficients:
•     Text to Bits Conversion: Convert the watermark text to a binary representation.
•     Hamming Encoding: Encode the binary bits using Hamming code for error detection
      and correction.
•     Embed Bits: Modify DCT coefficients based on the watermark bits.
8.    Apply Inverse DCT to Each Block.
9.    Combine the Channels and Convert Back to RGB.
10.   Save the watermarked image.
```

The output from each technique (Embedded Image) is sent to the next stage (Fake image generation) to generate fake images .. The fake image generation stage aims to generate counterfeit or synthetic images. It involves taking images that contain embedded watermarks and using them as inputs for StyleGAN2-ADA within the Google Colab environment. Google Colab, a cloud-based Jupyter notebook service by Google, provides a powerful platform to execute Python code, specifically catering to DL tasks while harnessing GPU resources. StyleGAN2-ADA in Colab uses pre-trained models to generate new images from watermarked ones. Google Colab provides essential GPU resources for StyleGAN2-ADA's computing needs. Free GPU access in Colab eliminates the need for costly hardware. Users can create synthetic images mimicking patterns from watermarked images. 40 images are saved for authenticity verification.

The image authenticity stage involves checking the authenticity of the images to decide whether the images are genuine or forgery. The suggested approach uses two methods: watermark analysis and image content analysis.

Watermark Analysis is the process of detecting, extracting, and analyzing watermarks in images, which is commonly used for authentication and integrity verification. The watermark's invisible text is detected and extracted from the images by employing two distinct methods due to the use of different embedding techniques (Algorithms 3 and 4) [23].

Algorithm 3: Watermark Extraction with Special Domain

Input: Watermarked Image.

Output: Text

1. Split the image into its color channels, namely, the Red (R), Green (G), and Blue (B) channels.
2. Store the pixel values of each channel individually into separate arrays.
3. Iterate through each array and subtract the value of either 0 or 1 (depending on the encoding scheme) from each pixel value.
4. Collect and interpret the resulting pixel values as characters to reconstruct the watermark text.

Algorithm 4: Watermark Extraction with Frequency Domain

Input: Quantized Watermarked Image

Output: Text

1. Load the watermarked image.
2. Convert the Image to YCrCb Color Space
3. Divide the Y Channel into 8x8 Blocks.
4. Apply DCT to Each Block.
5. Extract Watermark Bits from DCT Coefficients.
6. Decode Watermark Bits.
7. Combine the Watermark Bits into Text.
8. Obtain the extracted text.

These methods worked in tandem to reveal alterations induced by embedded watermarks, aiding in the differentiation between authentic and generated images. They evaluated the watermark's influence on the content and structure by extracting the watermark text and comparing it with known patterns. Methods were tested on the Flickr Faces High Quality (FFHQ) dataset using 12 real images with various characteristics. 12 fake images generated by StyleGAN2-ADA.

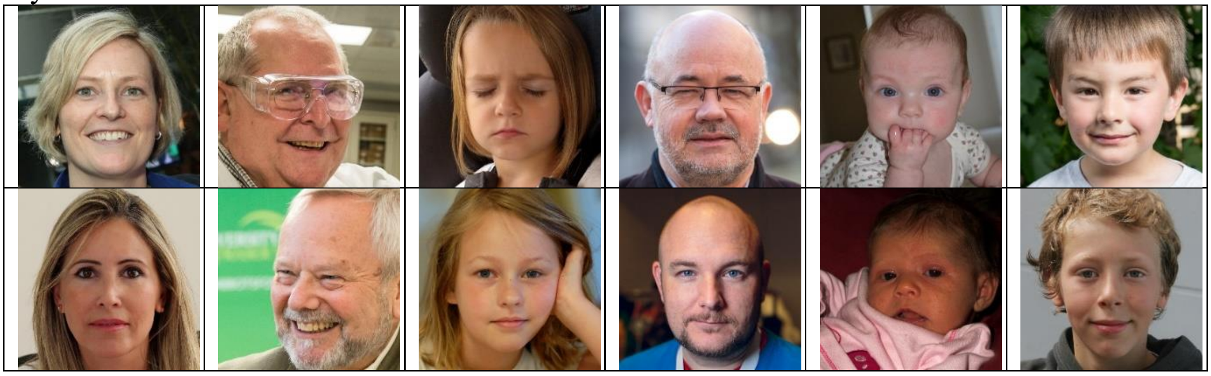


Figure 2: Real Images

Low-level features like color histograms quantify differences between real and fake watermarked versions based on pixel intensity. Metrics like PSNR, SSIM, MIS, and NCC identify potential manipulations. These metrics aid in evaluating similarity and detecting manipulated images. Together, they verify image authenticity and analyze content for fake image detection in digital forensics.

Experiment 2 Methodology

Image content analysis is crucial in digital forensics for detecting fake images by extracting and analyzing features and patterns. Techniques from digital forensics, computer vision, and machine learning are combined to detect fake images. ELA is used to detect image tampering by highlighting differences in compression levels. Perceptual image hashing generates content-based image hashes that are useful for detecting duplicate images and filtering inappropriate imagery [24]. Finally, CNNs are used to detect fake images. This model is popular in computer vision tasks for its effectiveness and availability of pre-trained weights. The network takes a 256*256 color image as input. Data augmentation creates additional image versions for better generalization. The network has three convolutional layers with different filter sizes and kernel sizes of 3x3. We added a max-pooling layer with a 2x2 filter and a stride of 2 after each layer. The rectified linear unit (ReLU) activation function is used to output the three convolutional layers. The stack of convolutional and max-pooling layers extracts the features from the input image, then these features are followed by a flatten layer for reshaping the output of the preceding layer into a one-dimensional vector, which is then fed into a dense layer with 128 neurons and ReLU activation and a second dense layer with 1 neuron to classify images into fake or real with a sigmoid activation function. The network was trained using 32 batch size and 10 epochs. Adaptive Moment Estimation (Adam) is used to optimize the weights in each filter. The model assesses the output to determine image authenticity. The CNN model was trained on the FFHQ dataset with 2000 images. The dataset includes real and fake images. Performance was evaluated using the same dataset as in Experiment 1. Precision, recall, F1-score, and accuracy were used for measurement. Equations 1, 2, 3, and 4 define the performance metrics [25].

$$Precision = \frac{TP}{TP+FB} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1 = 2 \times \frac{Precision+Recall}{Precision \times Recall} \tag{3}$$

$$Accuracy = \frac{TP+TN}{FP+FN+TP+TN} \tag{4}$$

Where:

- TP (True Positives): The number of correctly identified fake images.
- FP (False Positives): The number of images incorrectly classified as fake (but are real).
- FN (False Negatives): The number of fake images incorrectly classified as real.
- TN (True Negatives): The number of correctly identified real images.

Experiment 3 Methodology

The final stage integrates outcomes from two fake detection procedures: watermark and image content analysis. The goal is to consolidate findings for a more accurate determination of image authenticity, which is pivotal in legal proceedings. Amalgamating results achieve a comprehensive assessment of authenticity. When analyses align, confidence in the result significantly increases. The fusion of procedures provides a holistic perspective on genuineness. If all analyses agree on authenticity, confidence in the determination is higher. This consolidated evidence, from multiple detection techniques aligned in assessment, serves as compelling evidence in legal contexts, strengthening decision reliability on image content and authenticity.

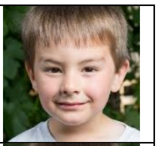
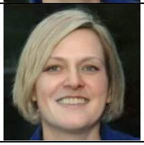
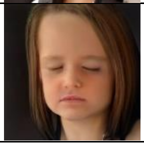
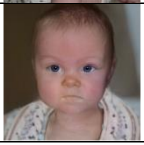
4. Results and Discussion

The following sections show the performance of the watermark image analysis and the content image analysis approaches, as well as the evaluation of the fusion approach.

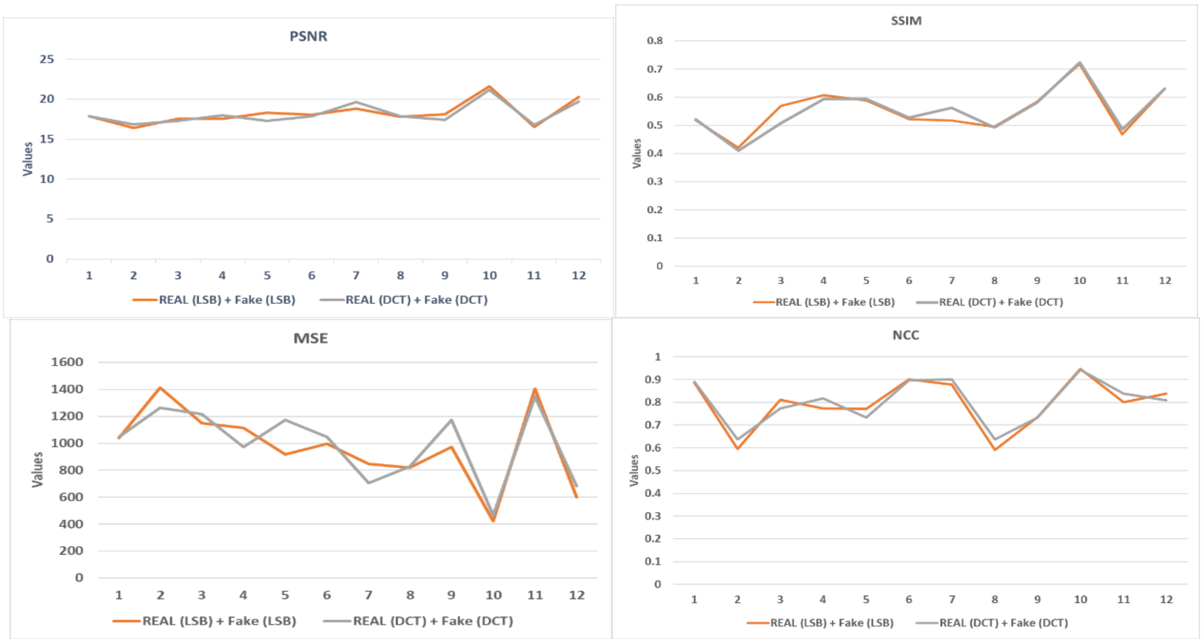
Experiment 1

The experiment investigated embedding invisible text into images using LSB and DCT techniques. The impact on image quality was assessed using the STYLGAN2-ADA algorithm. A Microsoft Visual Studio Python script was used for the experiment. 5,000 characters were embedded into each image. Results showed no significant differences between LSB and DCT techniques, as illustrated in Table 1.

Table 1: Comparison between Real Images and Fake Images using different embedding watermark techniques.

Real Images with LSB						
Fake Images with LSB						
Real Images with DCT						
Fake Images with DCT						

Objective image quality metrics (e.g. PSNR, SSIM, SIM, NCC, Histogram) compared real and fake images generated by STYLGAN2-ADA. Results show little difference in metric values, indicating that the choice of watermarking technique (LSB or DCT) did not impact fake image quality. The technique used to embed the watermark did not affect STYLGAN2-ADA's ability to generate realistic fake images resembling the originals. The text watermark proved efficient in detecting forged images, as the extracted text from the fakes was nonsensical regardless of the technique used. This is because STYLGAN2-ADA alters the image content, including the embedded text, making it difficult to detect the forgery despite the close visual similarity to the original image. MSE was considered the clearest metric for evaluating the quality of the generated fake images. The use of MSE as the primary metric for differentiating between the two image processing methods and the specific analysis of images 2, 4, 5, 7, and 8 due to changes in image features.



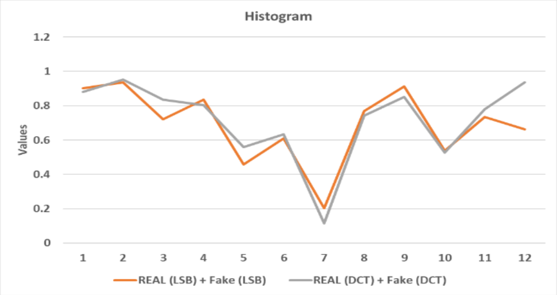


Figure 3: Objective Metrics to Compare between Real and Fake Images with Watermarking Techniques

The impact of watermarking on the clarity of fake images, with a more pronounced decrease when using the LSB technique, is shown in Figure 4. The effectiveness of watermarking in detecting fake images, even without the original source, and the use of retrieved corrupted text as proof of inauthenticity.

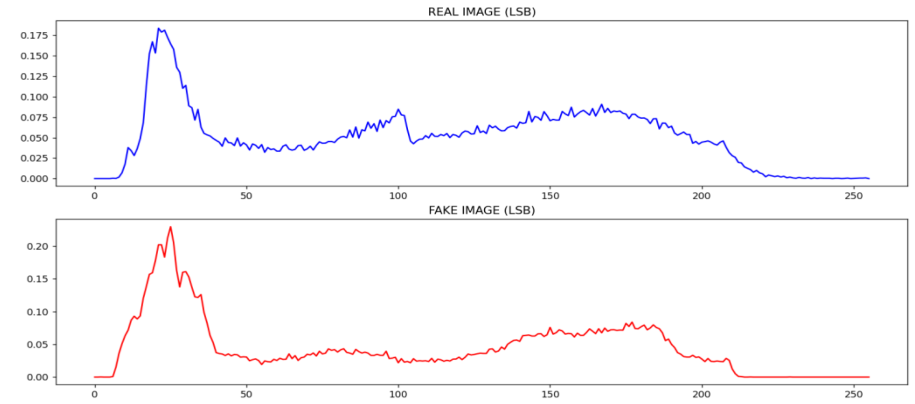
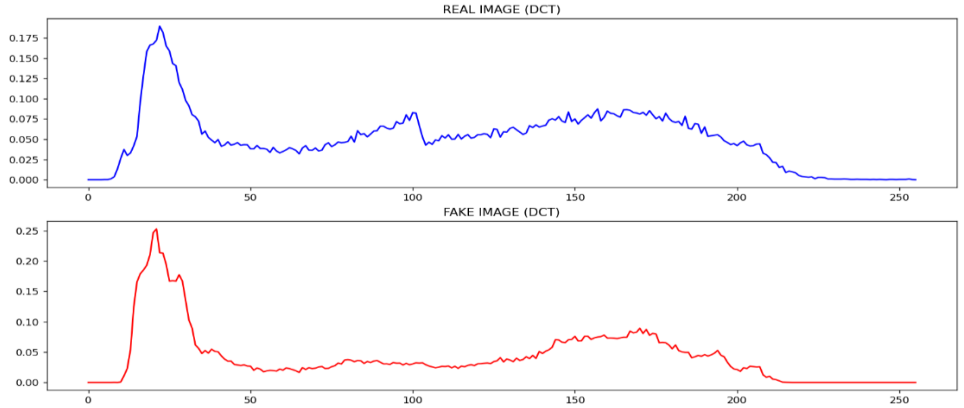
	Histograms Comparison
	 

Figure 4: Histogram Comparison

Experiment 2

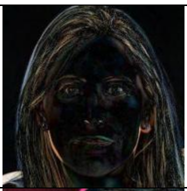
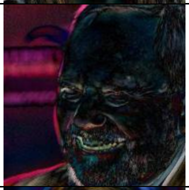
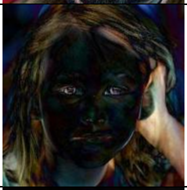
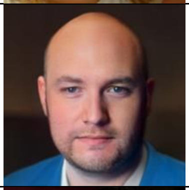
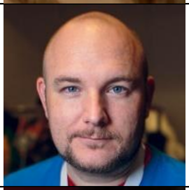
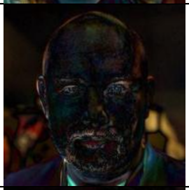
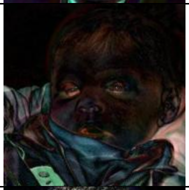
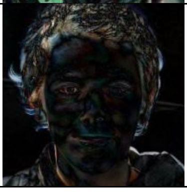
In this experiment, using Python code, ELA was applied to both real and manipulated images. The images were re-compressed at a fixed compression ratio, and the error levels were analyzed across different regions of each image. ELA detected altered regions. In ELA images, areas of consistent compression tend to exhibit similar error levels, usually appearing uniform. However, regions that have been edited or altered typically show different levels of error and may appear either brighter or darker. Brighter areas indicate higher error levels.

The ELA results for real images showed uniform error levels across the entire image, indicating no signs of post-compression manipulation. In contrast, the fake images exhibited varying error levels, particularly in manipulated regions, suggesting that these areas were edited. The findings suggest that the fake images were likely manipulated after their initial creation, as evidenced by the non-uniform error levels. The ELA results provide a clear visual distinction between real and manipulated images, making this technique valuable for detecting image forgery.

ELA proved to be a useful tool in distinguishing between real and manipulated images. The clear differences in error levels helped identify regions of interest, highlighting the potential of this technique in image forensics. When conducting forensic analysis, edited regions in an image will show distinct error levels compared to the rest of the image.

Regarding ELA, the original images were compared to the forged ones, and the results are shown in Table 2. This method is considered new in determining image authenticity because it efficiently identifies the areas that have been altered. However, the problem with this method is the requirement to have the original image to verify the authenticity of the target image. In the absence of the original image, ELA can still provide information about the possibility of image manipulations. Still, it becomes less accurate and effective than when the original image is available. When the image contains areas that have been subsequently modified (such as adding or removing elements or altering colors or details), these areas will exhibit a different error level than the rest. Without the original image, these areas can be compared based on the expected consistency in the error level across the image.

Table 2: ELA Results.

Real Image	Fake Image	ELA between Real and Fake Images
		
		
		
		
		
		

In detecting fake images, Perceptual Hashing provides a robust method to identify manipulated or altered images by comparing their visual 'fingerprints' with those of original images, even when the alterations are subtle. In this study, we applied Perceptual Hashing to a dataset of real and manipulated images. We used Perceptual Hashing to analyze a dataset of 80 real and modified photos. Each image was hashed using a particular algorithm or library, like pHash. The produced hashes were later analyzed to identify and point out the similarities and differences between actual and potentially fake (modified) images. This was done using the Hamming Distance technique, which measures the number of bits that differ between two hashes; therefore, a smaller Hamming Distance indicates greater similarity between the images, whereas a larger distance shows major differences. The results showed that perceptual hashing was extremely successful at detecting false visuals that had undergone subtle changes.
A smaller threshold value (T) typically leads to a lower collision probability, meaning fewer false positives when comparing images, as shown in Table 3. This is crucial for applications requiring high accuracy. The results appear in Table 4.

Table 3: Applying different threshold values.

Threshold Value	Accuracy	Incorrect number of predicted images
10	%88.5	7
9-8	%93.7	5
7-2	%97.5	2
1	%100	0

Table 4: Performance measures.

	Precision	Recall	F-measure	Support
Real	0.80	1.00	0.89	40
Fake	1.00	0.75	0.86	40
Accuracy			0.88	80

This means that the model correctly classified 88% of the images. Out of 8 images, it got 7 correct predictions.

- The model performs better on the Real class (recall of 100%) than on the Fake class (recall of 75%). It missed one fake image and classified it as real.
- The model has a high precision for both classes, meaning that it is highly confident and likely to be correct when it makes a prediction (whether real or fake).
- F1-scores for both classes are close (89% for real and 86% for fake), indicating balanced performance.
- The model is very good at identifying real images (perfect recall for real).
- It is also quite good at identifying fake images but can occasionally misclassify a fake image as real (as seen from the lower recall for fake).
- Given the overall accuracy of 88%, the model can be considered effective for this small dataset, though improvements could be made in identifying all fake images.
- The model is very good at identifying images as fake when they are truly fake (no false positives). This is critical in scenarios where you want to ensure that any flagged fake images are indeed fake.
- The model accurately detects fake images, which is critical for ensuring that flagged fake images are truly fake.
- Lower recall for fake class (75%): the model missed 1 fake image, classifying it as real, which may be problematic in forensic or security contexts.

- Balanced performance on real class: the model has high recall (100%) and reasonable precision (80%), successfully identifying real images but may misclassify some fakes.

Experiment 3

In our proposed framework, we utilize a hybrid approach that combines watermark analysis (using LSB and DCT techniques) with content-based image analysis methods (including Error Level Analysis (ELA), Perceptual Hashing, and Convolutional Neural Networks (CNN)). The integration of these two methodologies allows us to leverage their complementary strengths to enhance detection accuracy.

Watermarking techniques embed imperceptible signatures within images during creation. Our study employed LSB and DCT methods to embed hidden text watermarks. These watermarks serve as a reference for authenticity verification.

During detection, we check watermarks in images for integrity using ELA and compression artifact analysis. Perceptual hashing compares images based on visual features. If analysis methods conflict, we cross-validate. Confidence scores are based on watermark visibility and content metrics. The final decision is made by averaging high confidence scores.

If there is a significant discrepancy between scores (e.g., one method indicates authenticity while another indicates manipulation), we apply a heuristic decision rule that considers historical performance data for each method. Table 5 summarizes the three experiments used in our work.

Table 5: The three experiments' summaries.

Experiment	Description	Parameters Used
Experiment 1: Watermarking Methods	Evaluates the effectiveness of watermarking techniques (LSB and DCT) on real-life images to detect tampering.	**LSB Watermarking:** - Embedding Strength: 5,000 characters - Image Format: PNG - Image Size: 256x256 pixels (cropped and resized) - Conversion to Grayscale: Yes - Bit Conversion: Yes (text to bits) **DCT Watermarking:** - Color Space: YCrCb - Block Size: 8x8 pixels - Hamming Encoding: Yes (for error detection and correction)
Experiment 2: Content Image Analysis	Utilizes various techniques (including ELA, Perceptual Hashing, and CNN) to analyze image content for authenticity verification.	- ELA Method: Analyzed error levels across regions. - Perceptual Hashing: Used Hamming Distance for comparison. - CNN Architecture: Three convolutional layers. - Dataset Size: 80 images (real and manipulated) - Threshold Values for Accuracy: Varied from 1 to 10.
Experiment 3: Fusion Approach	Combines watermark analysis and content analysis techniques (including CNN) to improve overall detection performance.	- Watermark Analysis Method: LSB and DCT results. - Content Analysis Techniques: ELA results, Perceptual Hashing comparisons, and CNN outputs. - Decision Fusion Method: Weighted confidence scores based on performance metrics from both analyses.

5. Conclusion

This study presents a good approach in digital forensics for identifying counterfeit images by using GANs (Generative Adversarial Networks) through the style GAN2-ADA algorithm, combined with the concealed watermark text methods and image content analysis. Striving to preserve the credibility of visual information in an era where distinguishing between reality and deception is becoming harder and more complex. The model suggested combines the results of two different processes for detecting false content: image content analysis and watermark analysis. Using this technique, GAN silently embed an invisible watermark onto images, acting as a unique identifier for genuine verification, utilizing GANs' ability to generate realistic fake images. It uses both spatial and frequency domain approaches to embed watermarks and identify imperfections. Image content analysis compares the original and suspicious photos and looks for differences using CNN, visual hashing, and ELA. Combining watermarking, classification, and image content analysis results in an extensive and dependable method for identifying invalid photos. The model's results demonstrate that this tactic may prevent the spread of modified photos, uphold photojournalistic integrity, and validate digital evidence in cases.

In the future, one can explore transfer learning approaches using pre-trained models on large datasets to improve performance on smaller datasets commonly found in digital forensics. Conduct field studies in collaboration with law enforcement or media organizations to evaluate the effectiveness of proposed methods in real-world scenarios. Also, longitudinal studies should be conducted to track the evolution of image manipulation techniques over time, allowing for adaptive detection strategies that evolve alongside new threats.

References

[1] V. V. V. N. S. Vamsi, S. S. Shet, S. S. M. Reddy, S. S. Rose, S. R. Shetty, S. Sathvika, et al., "Deepfake detection in digital media forensics," *Global Transitions Proceedings*, vol. 3, pp. 74-79, 2022.

[2] L. Nataraj, T. M. Mohammed, S. Chandrasekaran, A. Flenner, J. H. Bappy, A. K. Roy-Chowdhury, et al., "Detecting GAN generated fake images using co-occurrence matrices," arXiv preprint arXiv:1903.06836, 2019. doi:10.2352/ISSN.2470-1173.2019.5.MWSF-532

[3] T. Osakabe, M. Tanaka, Y. Kinoshita, and H. Kiya, "CycleGAN without checkerboard artifacts for counter-forensics of fake-image detection," *International Workshop on Advanced Imaging Technology 2021,* pp. 51-55, 2021. doi:10.1117/12.2590977

[4] O. Mayer and M. C. Stamm, "Exposing fake images with forensic similarity graphs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 1049-1064, 2020. doi:10.1109/JSTSP.2020.3001516

[5] Y. Guo, X. Cao, W. Zhang, and R. Wang, "Fake colorized image detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1932-1944, 2018. doi:10.1109/TIFS.2018.2806926

[6] T. M. Ishwarya and K. N. Durai, "Detection of face mask using convolutional neural network," *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 2008-2012, 2022. doi:10.1109/ICACCS54159.2022.9785054

[7] A. Ferreira, E. Nowroozi, and M. Barni, "Vipprint: Validating synthetic image detection and source linking methods on a large scale dataset of printed documents," *Journal of Imaging*, vol. 7, no. 50, 2021. doi:10.3390/jimaging7030050

[8] W. H. Abdulsalam, S. Mashhadani, S. S. Hussein, and A. A. Hashim, "Artificial Intelligence Techniques to Identify Individuals through Palm Image Recognition," *International Journal of Mathematics and Computer Science*, vol. 20, pp. 165-171, 2025. doi:10.69793/ijmcs/01.2025/abdulsalam

[9] S. Mashhadani, W. H. Abdulsalam, O. A. Hassen, and S.-M. Darwish, "Fusion of Type-2 Neutrosophic Similarity Measure in Signatures Verification Systems: A New Forensic Document Analysis Paradigm," *Intelligent Automation and Soft Computing*, vol. 39, no. 5, pp. 805-828, 2024, doi:10.32604/iasc.2024.054611

[10] R. G. Mani, R. Parthasarathy, S. Eswaran, and P. Honnavalli, "A survey on digital image forensics: Metadata and image forgeries," *WAC-2022: Workshop on Applied Computing*, pp. 22-55, 2022.

[11] Z. M. J. Kubba and W. A. Shukur, "An enhanced LED cipher algorithm performance for data security in IoT systems," *The Second International Conference on Emerging Technology Trends in Internet of Things and Computing*, vol. 3015, no. 1, 2023. doi:10.1063/5.0188272

[12] W. A. Shukur, Z. M. J. Kubba, and S. S. Ahmed, "Novel Standard Polynomial as New Mathematical Basis for Digital Information Encryption Process," *Advances in Decision Sciences*, vol. 27, pp. 72-85, 2023. doi:10.47654/v27y2023i3p72-85

[13] K. Karampidis, E. Kavallieratou, and G. Papadourakis, "A review of image steganalysis techniques for digital forensics," *Journal of Information Security and Applications*, vol. 40, pp. 217-235, 2018. doi:10.1016/j.jisa.2018.04.005

[14] Z. M. Nabat, S. A. Shnain, B. J. Al-Khafaji, and M. A. Salih, "Steganography of image based on random key generation and XOR LSB substitution," *in AIP Conference Proceedings*, p. 020022, 2025. doi.org/10.1063/5.0289728

[15] A. A. R. Bsoul and Y. Alshboul, "Integrating Convolutional Neural Networks with a Firefly Algorithm for Enhanced Digital Image Forensics," *AI*, vol. 6, p. 321, 2025. doi.org/10.3390/ai6120321

[16] F. K. Al Jibory, O. A. Mohammed, and M. S. H. Al Tamimi, "Age estimation utilizing deep learning Convolutional Neural Network," *International Journal on Technical and Physical Problems of Engineering*, vol. 14, pp. 219-24, 2022.

[17] M. M. Shwaysh, S. Alani, M. A. Saad, and T. A. Abdulhussein, "Image Encryption and Steganography Method Based on AES Algorithm and Secret Sharing Algorithm," *Ingenierie des Systemes d'Information*, vol. 29, p. 705, 2024. doi.org/10.18280/isi.290232.

[18] W. H. Abdulsalam, R. S. Alhamdani, and M. N. Abdullah, "Emotion recognition system based on hybrid techniques," *International Journal of Machine Learning and Computing*, vol. 9, no. 4, pp. 490-495, 2019. doi:10.18178/ijmlc.2019.9.4.831

[19] K. Remya Revi, K. R. Vidya, and M. Wilscy, "Detection of deepfake images created using generative adversarial networks: A review," Second International Conference on Networks and Advances in Computational Technologies, pp. 25-35, 2021. doi:10.1007/978-3-030-49500-8_3

[20] P. Yang, D. Baracchi, R. Ni, Y. Zhao, F. Argenti, and A. Piva, "A survey of deep learning-based source image forensics," *Journal of Imaging*, vol. 6, 2020. doi:10.3390/jimaging6030009.

[21] Jayapandiyan, Jagan Raj, C. Kavitha, and K. Sakthivel, "Enhanced least significant bit replacement algorithm in spatial domain of steganography using character sequence optimization," *IEEE Access,* vol. 8, pp. 136537-136545, 2020. doi: org/10.1109/ACCESS.2020.3009234

[22] Yuan, Zihan, Qingtang Su, Decheng Liu, and Xueting Zhang, "A blind image watermarking scheme combining spatial domain and frequency domain," *The visual computer,* vol. 37, pp. 1867-1881, 2021. doi:org/10.1007/s00371-020-01945-y

[23] Begum M, Uddin MS. "Digital image watermarking techniques: a review," *Information,* vol. 11, no. 2, 2020. doi.org/10.3390/info11020110

[24] P. Samanta and S. Jain, "Analysis of perceptual hashing algorithms in image manipulation detection," *Procedia Computer Science*, vol. 185, pp. 203-212, 2021. doi:10.1016/j.procs.2021.05.021

[25] Ban Hassan, Majeed, Wisal Hashim Abdulsalam, Zainab Hazim Ibrahim, Rasha H Ali, and Shahlaa Mashhadani, "Digital Intelligence for University Students Using Artificial Intelligence Techniques," *International Journal of Computing and Digital Systems*, vol. 17, no. 1, pp. 1-10, 2025, doi:org/10.12785/ijcds/1571029446.