Evaluation of Spiral Pattern Watermarking Scheme for Common Attacks to Social Media Images

Tiew Boon Li[1], Jasni Mohamad Zain[2], Dr. Syifak Izhar Hisham[3], Alya Afikah Usop[4]
Faculty of Computing, Universiti Malaysia Pahang, Lebuhraya Tun Razak, Pahang[1, 3]
Institute for Big Data Analytics and Artificial Intelligence, Kompleks Al-Khawarizmi, UiTM, Shah Alam, Malaysia[2]
Faculty of Computing, Universiti Malaysia Pahang, Pekan, Pahang[4]

***Abstract*—The 21st century might be considered the "boom" period for social networking due to the fast expansion of social media use. In terms of user privacy and security regulations, a plethora of new requirements, issues, and concerns have arisen due to the proliferation of social media. With the increase in social media use, images on social media are often modified or fabricated for certain purposes. Therefore, this work implements and evaluates the SPIRAL-LSB algorithm for common attacks for social media images. Image compression was also discussed as images published to social media platforms was often compressed. An analysis was performed to assess the algorithm's output on social media images. The experiments were carried out prior to and after uploading to the Instagram platform. The dataset was subjected to image splicing, copy-move, cut-and-paste, text insertion, and 3D-sticker insertion attacks. The outcome of SPIRAL-LSB was effective for text insertion attacks solely. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) were selected as the experiment's metrics. The average PSNR value is 63.25, and the SSIM value is 0.99964, both of which are regarded high. This indicates that the watermark has not degraded the quality of the images. This work was designed for usage on social media for intellectual property reasons and may be used to validate the validity of social media images and prevent issues with image integrity, such as image manipulation.**

Keywords*—Spiral pattern; fragile watermarking; social media; LSB substitution

I. Introduction

The application of social media is rapidly intensifying, and the twenty-first century may be defined as the "boom" time for social networking. According to Smart Insights data, there were approximately 3.484 billion social media users in February 2019. The Smart Insight survey reported the number of social media users is increasing by 9% every year, and this trend is expected to continue [1]. Currently, the social media users symbolize 45% of the worldwide population [2]. The most frequent users of social media are digital natives, a group of people who were born or grew up in the digital age and are familiar with numerous technologies and systems, and the Millennial Generation, individuals who became adults around the turn of the twenty- first century.

Moreover, according to [3], the usage and sharing of information through the internet is an inherent or intrinsic element of university students' lives. The study's results indicate that students often use Facebook to share information. Numerous individuals disclose their personal information without thinking of the consequences. Consequently, social media platforms have developed into a vast repository of sensitive data. Users are more receptive to friend invitations and trust goods sent to them by friends [4].

The move toward visual social media is being pushed in part by changes in social media user behaviors as a result of the enhanced mobile internet experience. Due to the widespread use of advanced software applications such as Picasa and Photoshop, image manipulations have become a fairly popular and effortless action for everyone. Edited images are often aesthetically appealing and difficult to differentiate from unaltered ones. There is a growing tendency toward the use of modified images in every aspect of our daily lives, such as news reporting, blogging, and advertising [5]. This often leads to user deception [6] which has the potential to influence and manipulate public opinion, ranging from teens' self-esteem and personal health choices to public opinion in significant political areas.

Although manipulated images are often uncovered, it may take weeks, and by that time, millions of people's opinions have already been influenced. This may raise severe concerns about the trustworthiness of digital multimedia, since it puts questions on the face value of the information we receive on a regular basis through the Internet [7]. This issue is getting more severe, presenting major difficulties to society. Revolution of Internet and technology enables pirates to unlawfully utilize the features to manipulate images [8]. Thus, the necessity for digital media authentication techniques becomes vital to ensuring that work is not tampered with, particularly in crucial circumstances such as social media politics, medical safety, internet banking, military data transmission, and forensic investigations.

In disciplines such as forensics, medical imaging, and military and industrial images, the integrity of a digital image is critical [9]. Digital watermarking is considered a technological category in dealing with integrity issues [10]. Hence, to preserve social media images and identify ownership, digital watermarking is essential. Without watermarks, images on social media are vulnerable to theft and illegal use [11]. In theory, digital watermarking can distinguish between various sorts of third- party manipulations and attacks.

A. Integrity and Authentication of Digital Images

In between the techniques for securing digital data, digital watermarking has grown in popularity among academicians and users due to its variety and ability to retain the integrity

and authenticity of digital images. The term "image authentication" refers to the process of determining the legitimacy of digital images. Among the methods for establishing image authenticity are location of tampering. As mentioned in the previous section, digital watermarking may potentially discriminate between different types of manipulations and assaults by a third party. Manipulations in this instance include those that are permitted and those that are not permitted [12].

There are three techniques to watermarking that includes fragile watermarking, robust watermarking, and semi-fragile watermarking, which combine fragile and robust aspects. Watermarking images is critical for preserving personal data privacy and avoiding image tampering [10]. In general, an image authentication technique is composed of two stages: embedding and validation. The embedding stage embeds the authentication data in an image and stores it as proof of the image's validity; the validation stage compares two images: one evaluated for the watermarked image, and another extracted from the watermarked image and determines whether the image has been modified or not [13].

Authentication through fragile watermarking is performed by embedding a watermark into the image, which is quickly altered or destroyed when the watermarked image is manipulated or attacked. When compared to the image's real content, the presence or absence of the watermark is identified [12]. Several prominent strategies allow for the localization and recovery of changed regions in a block-wise manner. While embedding, certain techniques may provide metadata about the image. In contrast, systems based on robust watermarking assume that a good watermark is impervious to image manipulations.

Digital watermarking, among current approaches and owing to its exceptional qualities, is an efficient option for protecting multimedia data in a variety of industries. The primary benefit of digital watermarking is that the authentication data is included directly in the image data. The authentication information is preserved, even if the watermarked image is converted to a different format and the retrieval procedure is described as simpler and less complex [14]. Table I shows the key contrasts between these three notions namely cryptography, steganography, and watermarking. These three methods are commonly used as data security techniques. Fig. 1 shows the data security techniques.

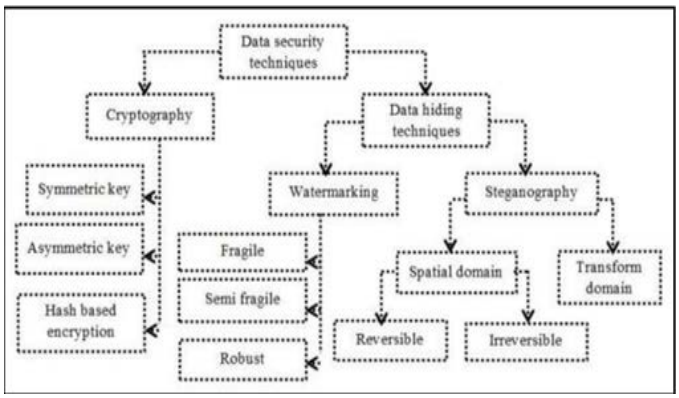


Fig. 1. Data Security Techniques [14].

TABLE I. COMPARISON OF CRYPTOGRAPHY, WATERMARKING AND STEGANOGRAPHY [14]

Criterion	Cryptography	Watermarking	Steganography
Objective	Encrypted communication	Content authentication and copyright preservation	Covert communication
Authentication	Yes	Yes	No
Cover selection	Not required	Usually image, audio, or video	Any digital object
Key	Mandatory	Optional	Optional
Attacks	Cryptanalysis attacks; Ciphertext only attacks; Known-plaintext attack; Chosen-plaintext attack; Brute-force attack; Man-in the-middle attack; Birthday attack; Timing attack; Dictionary attack.	Image processing attacks; Salt and pepper noise; Cropping attack; Rotation attack; Sharpening attack; JPEG attack; Median filtering attack; Quantization; Temporal modification.	Steganalysis attacks; Regular and singular analysis; Pixel difference histogram attack; Chi-square attack; Sample pair analysis.
Robustness	Not required	Should be high	Should be high
HC	Not required	Should be high	Should be high
Imperceptibility	Not required	Should be high	Should be high
Visibility	Always visible	Depending upon the type of watermarking, it can be visible or invisible	Always visible
Output	Encrypted text	Watermarked object	Camouflage object
Merits	It offers both authentication and integrity, along with confidentiality	It offers both authentication and integrity, along with confidentiality	None apart from the sender and receiver can suspect the existence of the communication
Demerits	The communication is visible to the outsider	HC is usually low	Steganography itself alone cannot provide authentication and integrity
Purpose is lost	If the communicating message is decrypted	If the watermark is abolished or heavily tampered	If the attacker knows communication
Origin	Very ancient	Modern era	Very ancient

The remainder of this paper is organized as follows: Section 2: Literature Review, Section 3: Methodology, Section 4: Results and Discussion, Section 5: Conclusion, Section 6: Acknowledgement and Section 7: References.

II. LITERATURE REVIEW

A. Related Works

An overview of fragile watermarking systems for image authentication is presented by [15]. The limited embedding capability and amount of tampering are two major challenges

that motivate study in this field. This review covers the overall framework of the fragile watermarking system, as well as the many types of assaults and parameters used to evaluate the methods. The researchers will be able to quickly analyze current achievements in this field by using comparative analysis and quantitative comparisons of fundamental schemes and their variants with enhancements.

The authors [16] propose a secure fragile image watermarking system that is used to identify image content alteration or manipulation. The proposed approach consists of two steps: computing a secure authentication code/watermark bit from some of each pixel's most significant bits, and then hiding the watermark bit in the least significant bit (LSB) of each pixel using a recommended watermark embedding procedure. On a series of grayscale images, the proposed watermarking method is evaluated, and the watermarked image's quality is shown.

The authors [17] present a dual watermarking technique capable of integrating authentication, copyright protection, and image recovery functionalities into a single cover image. The robust scheme protects against copyright infringement by utilizing a single watermark in the discrete cosine transform (DCT) domain, whereas the fragile scheme protects against copyright infringement by utilizing two self- embedding watermarks in a spatial domain for authenticating and restoring digital image content.

The authors [18] presented a new technique for copyright protection, data security and content authentication of multimedia images. The authentication of the content has been ensured by embedding a fragile watermark in the spatial domain while copyright protection has been taken care of utilizing a robust watermark. The fragile watermark embedding makes the system capable of tamper detection and localization with average value more than 45% for all signal processing and geometric attacks. The average Peak-Signal- to-Noise Ratio (PSNR) achieved for both schemes are greater than 41 dB.

The author [19] developed a unique spiral numbering pattern for fragile digital watermarking schemes. The developed scheme is designed to achieve a good numbering pattern, exact detection, and image recovery. The limitations of the proposed scheme are works on gray-scale images only and square images.

To address the identified gap in the watermarking literature, the majority of studies have been conducted on medical images; however, there are a few studies that have been conducted on the security of social media images via digital watermarking, and the aforementioned studies have their own limitations and weaknesses. Thus, this work implemented and evaluated a fragile watermarking method on social media images. Initially, this algorithm was shown to function for medical images but has not been demonstrated to work for social media images. Thus, our effort adds to the security and integrity of images shared on social media platforms such as Instagram.

B. Popular Social Media Platforms

- Facebook: Facebook is a large social networking website where users may share comments, photos, and links to news or other relevant items on the web, as well as live chat and watch reels. Shared information may be made publicly available or restricted to a small group of friends or family members, or to a single individual. Since its inception on February 4, 2004, Facebook has grown to over 1.59 billion monthly active users, making it one of the finest platforms for connecting people from all over the globe.

- Twitter: Twitter is ranked as one of the top social networks in the world by active users. Twitter has 192 million marketable daily active users and gains 5 million daily users in the fourth quarter of 2020 [20]. Twitter gains 5 million daily users in Q4, Projects 20 Twitter is a popular social media site because it is personal and rapid. Twitter combines instant messaging, blogging, and texting, but with brevity and mass appeal. Most people nowadays have Twitter accounts including celebrities who use Twitter to engage with followers.

- Instagram: Instagram is one of the most popular social media platforms in the modern day. Without Instagram, it is difficult to run an effective social media marketing strategy. As an image and video-centric social network, Instagram gained popularity due to its easy filter tool, which can instantly transform any shot into a high-quality one. Live video, Instagram TV-IGTV, geotagged posts, hashtags, stories, and advertisements all appear as attractive features for users. Hence, the site has around 400 million active users and was acquired by Meta in 2012. Most people utilized Instagram to share information on travel, fashion, nutrition, and craftsmanship.

- WhatsApp: WhatsApp is a cross-platform instant messaging application available on smartphones, tablets, and personal computers. This program requires an Internet connection in order to transmit photos, text, documents, audio, and video messages to other users who have installed the app on users' devices. WhatsApp Inc. was founded in January 2010 and was acquired by Meta on February 19, 2004, for about $19.3 billion. Today, over a billion people use the internet to communicate with their friends, families, and even customers.

- Snapchat: Despite the competition from other social media platforms, Snapchat continues to be one of the most popular social media platforms today, especially among younger users. Indeed, in 2021, Snapchat had approximately 428 million users worldwide. Snapchat initially used it for private image sharing, video, and messaging, as well as generating caricatures like Bitmoji characters and sharing a chronological story with users' followers.

- Reddit: Reddit is a community-driven news website where users may produce and share content. The reason users of Reddit are attracted to the site is the promise of high-quality material. Reddit members are very active and often publish something fresh and intriguing. Reddit was one of the most popular mobile social applications in the United States as of June 2021, with around 48 million monthly active users.

C. Image Compression on Social Media

Working with larger photos with a higher bit depth, the images become too enormous to send over a regular network connection. To show an image in a fair length of time and utilize a reasonable amount of space to retain the image, approaches to minimize the image's file size must be used. These approaches analyze and compress visual data using mathematical algorithms, resulting in reduced file sizes. This is known as compression.

There are two types of image compression methods: lossy and lossless. Both systems conserve storage space, but the strategies used are different. Lossless compression expresses data in mathematical formulae while retaining all the original image's information. The integrity of the original image is preserved, and the decompressed image output is bit-for-bit identical to the original image input [21].

Lossy compression shrinks files by removing extra image data from the original image. It eliminates features that are too fine for the human eye to distinguish, resulting in close approximations of the original image, but not a perfect reproduction [21]. One of the most apparent advantages of lossy compression is that it results in a much lower file size compared to lossless compression, but at the expense of quality. With lossy compression, it is necessary to establish a compromise between file size and image quality. As shown in Fig. 2, with 50 percent compression, we reduced the size of the image file by 90 percent. With a compression ratio of 80%, we were able to reduce the image file size by 95%.

Lossless compression, on the other hand, is the process of reducing the size of an image without compromising its quality. Typically, JPEG and PNG files are stripped of unnecessary information. Lossless image formats include RAW, BMP, GIF, and PNG. With small reductions in image file sizes, there is no loss of image quality. Fig. 3 depicted the original and lossless compressed image.

Huge volumes of fresh data are constantly posted to Instagram's servers as a result of the millions of new posts submitted daily. The problem might soon spiral out of control if terabytes of data are uploaded every day. Instagram compresses both image and video postings to decrease server strain and maintain a steady flow of content. The user experience is also a factor in the compression. Some large videos and images would take a long time to upload if compression were not available. Users may be dissuaded from uploading further data if there are lengthy wait periods. In turn, this would result in decreased Instagram traffic and user engagement. Instagram has effectively avoided this problem, whether on purpose or not, by imposing rigorous limits and limitations on image sizes.

Fig. 2. Degree of Lossy Compression [22].

Fig. 3. Original Image and Lossless Compressed Image [22].

The issue arises due to Instagram's excessive JPEG compression of uploaded and shared images. JPEG employs lossy compression, which discards data, increasing the likelihood that watermarked data may be discarded. When users upload JPEGs to Instagram, they are compressed again, but by Instagram. In essence, users are double the compression and sacrificing quality. PNG uses lossless compression and hence should be less impacted by Instagram's. Uploading images in PNG format is advised to maintain a small size and good quality of images.

D. Common Attacks on Social Media Images

Digital images may be manipulated or attacked to deceive by changing some of the image's critical information. These attacks can be performed on social media images and lead to negative consequences such as financial loss, business fraud, defamation and to serious extent, cybercrime proceedings. These alterations are extremely destructive to some critical images, such as military and medical images, and such images should be preserved. The authors [23] categorized image forgery techniques into two basic approaches: active and passive, as seen in Fig. 4.

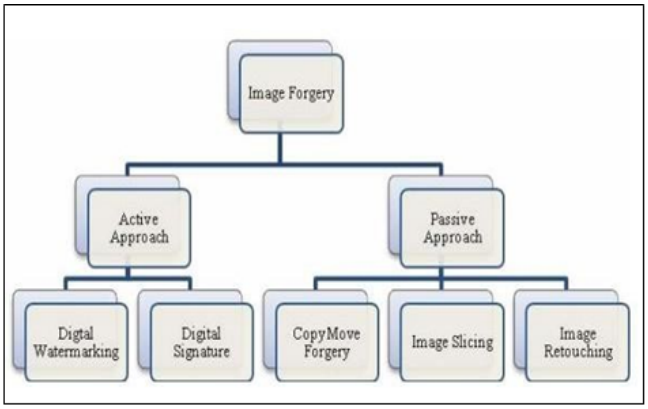


Fig. 4. Image Forgery Techniques.

1) Image splicing: Image splicing to generate counterfeit photos is more aggressive than image editing. Image splicing is a fundamentally basic procedure that may be done as crops and pastes regions from the same or other sources. This approach refers to a paste-up done by gluing together images utilizing digital tools available such as Photoshop.

In image splicing method there is composition of two or more photos, which are merged to generate a false image. Examples include some notorious news reporting situations involving the use of falsified images. Fig. 5 illustrates how to generate a forge image; by transferring a spliced piece from the source image into a target image, it creates a composite image of scenery which is a forge image.

The authors [24] describe image splicing as a collage created by adhering photographic images together. Image splicing is a method that combines two or more images to generate a new fictitious image. The image splicing technique is more aggressive than the resampling technique [24]. It is often followed by post processing such as blurring, compression, and scaling. It is often employed as the first stage in photomontage, a technique that is very popular in digital image content modification.

The authors [25] identified an image splicing approach that is based on image texture analysis, which defines image portions based on their texture richness. The texture content of an image is used to describe it in this manner. The modified image created by splicing might be utilized in news stories, photography contests, or as major evidence in academic papers, which could have a decisive impact.

2) Copy move attack: The copy move forgery is one of the commonly utilized forms of image manipulation method. In this approach, one has to cover a section of the image in order to add or delete information. In a copy-move attack, the objective is to disguise anything in the original image with some other section of the same image. The example of copy-move type is as shown in Fig. 6 when a troop of soldiers are cloned to cover George Bush.

The authors [24] claimed that copy move attack is when a portion of an image is copied and pasted into different locations within the same image to conceal information or change the meaning of the image. The digital image copy-move forgery technique involves the repetition ozone or more areas at various positions inside the same image. Frequently, duplicated portions are extended, shrunk, or rotated to increase the convincingness of forgeries, making it more difficult to identify forgeries.

3) Image retouching: Previously, retouched images were intended for magazine covers and mostly used on celebrities. The advancement of technology has increased the ease with which images may be retouched, resulting in a rise in over-perfect images. For example, Zendaya has taken to Instagram to criticize publications for retouching magazine figures as seen on Fig. 7.

Most alarming consequence of image retouching is the booming of selfie culture, which promotes a society preoccupied with money, beauty, power, and fame. Photoshop and Beauty Camera paving the way for unattainable beauty standards and are thereby contributing to the rising pandemic of body dysmorphia and mental health problems among today's youth. The image is not drastically altered during image retouching, but some characteristics of the image are enhanced or diminished, a technique that is quite common in the majority of photo editing software. In most image magazines, there is a need for image attractiveness, which results in the enhancement of some aspects of an image, oblivious to the fact that such approach is illegal.

4) Meme manipulation: The term "meme" derives from the Greek "mimesis," which refers to the way art imitates life [26]. Memes have been used as a weapon in cultural battles for more than a decade. Memes are more convincing than most people believe. On a social media timeline, a well-placed meme might lead down a rabbit hole of radicalization, misinformation, and extremism. In this scenario, Internet Memes stepped in as a compelling tool for users to express themselves in the ironic format, which often combines visual and text materials. Fig. 8 shows an example of a political meme between North Korea and America.

The authors [27] define Internet memes as artifacts of participatory digital culture, an excellent description of the functional purpose. Memes have the capacity to be made, utilized, spread, and remixed by anybody with Internet connection creating previously unimaginable opportunities for engagement in social and political concerns. To date, research on memes has been concerned with their contribution to the expression of political ideas and of subcultural identity [28].

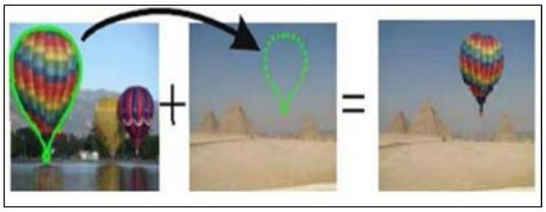


Fig. 5. Image Splicing.

Fig. 6. Copy Move Forgery Image.

Fig. 7. Zendaya Retouched Image.

Fig. 8. Political Meme.

III. Methodology

In the image authentication phase, collected sample images are input into the algorithms and output are obtained. Data collection starts the flow of the work. 15 original colored-images are acquired which acts as the host image. Host image was fed in the algorithm to be embedded with a watermark. The authentication watermark was a 2-bit authentication watermark, intended to compare the intensity (v) and the parity bit (p) for the detection of tamper in the colored-image. Following that, the host image undergoes block division to produce an image block (in pixels) using block numbering in a spiral pattern. After the embedding process, a watermarked image is produced.

For the purpose of testing, the watermarked image was manipulated with five different attacks, namely, image splicing, copy-move forgery, cut-and-paste, sticker insertion and text insertion. These five attacks are the most common attacks performed on social media images. The 15 sample images were manipulated with each type of attack, thus producing 75 attacked images as the input to the algorithm. To depict the image compression influence on social media images, the image authentication process was performed twice, first prior to the upload into social media and second after uploaded into social media.

The functional block diagram for watermark numbering, mapping, generation, and embedding was shown in Fig. 9. The technique in numbering is in a spiral manner. The following algorithms describe how the 2-tuple watermark of each sub-block was generated and embedded, which adapted from [19]:

1) Set the LSB of each pixel within the block of B to zero.

2) Calculate the average intensity of the block, AvgB and each of its sub-blocks, AvgBs, respectively.

3) Generate the authentication watermark, v, of each sub-block. V is 1 if the AvgBs is bigger than AvgB or 0 if otherwise,

4) Generate the parity check bit, p of each sub-block. P is 1 if the parity number is odd, and 0 if otherwise.

5) Obtain the original image, A, from the mapping sequence done at the first phase.

6) Compute the average intensity of each sub-block again within A, AvgAs.

7) Embed the 2-tuple watermark (v, p) each in one LSB of each pixel in Bs.

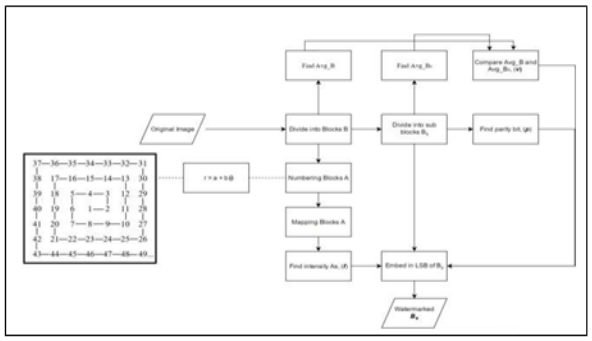


Fig. 9. Embedding Process [19].

In the SPIRAL-LSB scheme, two levels of detection phase were developed to guarantee no missing tamper when detecting. The first level would examine the parity bits and values of the average intensity in the sub-blocks, while the second level would examine the parity bits and values of the average intensity in the blocks containing the sub-blocks examined on the first level. This is done to ensure a high detection rate.

The experimental images were initially separated into non-overlapping 8 by 8-pixel blocks, similar to the watermarking embedding procedure. For each Br block, the LSBs of each Br pixel were set to zero and its average intensity, designated by Avg_Br, was computed. Then, a two- level detection was conducted. The procedure of hierarchical tamper detection scheme from [29, 30] is outlined below:

- Level 1 detection: For each 4 × 4-pixel sub-block Brs inside the block Br, do the following operations:

1) Extract v and p from Brs.

2) Set the LSBs of each pixel within each Brs to zero and compute the average intensity for each sub-block Brs, denoted as avg_Brs.

3) Set the algebraic relation v'=1 if avg_Brs >= avg_Br, otherwise, set it to 0.

4) Calculate the total number of 1s in avg_Brs and denote it as Ps.

5) Set the parity check bit p' of Brs to 1 if Ps is even, otherwise, set it to 0.

6) Compare p' with p and compare v' with v. If unequal, mark Brs as tampered and complete the detection for Brs; otherwise mark it as valid.

- Level 2 detection: For each valid 8x8 pixel block Br, do the following operations:

1) Search the block number of block C, where block C is the one in which the intensity feature of block Br is embedded.

2) Locate block C.

3) If block C is marked tampered, assume block Br is valid and complete the test.

4) If block C is valid, perform the following steps:

a) Get the 7-bit intensity of each Brs by extracting the LSBs from each pixel in the corresponding block within block C, padding one zero to the end to make an 8-bit value.

b) Compare with avg_Brs and mark Br tampered if they are different.

IV. Results and Discussion

After The samples to test the SPIRAL-LSB scheme were in PNG and JPG format with RGB colored type. The images were in square-sizes. From the algorithm applied, our result showed that embedding scheme with the block spiraling and starting the numbering in the middle would produce significant PSNR values, which as a whole, we can say all were above 55 dB, with average of 65.09 dB reported from the output data of 15 samples.

The highest value was 67.5 dB and lowest was 58.98 dB. Fig. 10 depicted the graph of the recorded PSNR values. Moreover, the SSIM value produced a correlation average value of 0.99964 which we regarded as very high. The produced SSIM value corresponds to one, indicating that the watermarked image closely resembles the original. The highest and lowest values were 0.9992 and 0.9998, respectively. Fig. 11 shows the graph of the recorded SSIM value.

A. Text Insertion Attack

In Fig. 12, a text "Vaccinated!" was inserted on the image (a) to produce tampered image (b). After acquiring the tampered image, it was tested prior to uploading it to Instagram. Figure (c) is the result before uploading while figure (d) shows the result after uploading. The tampered region is detected in red color. The tamper was detected and marked it in red, as shown in Fig. 12(d).

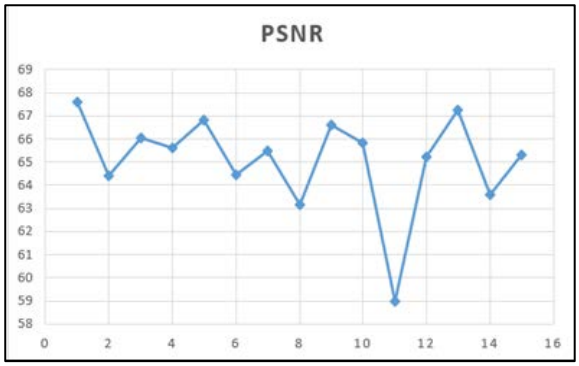


Fig. 10. PSNR Value.

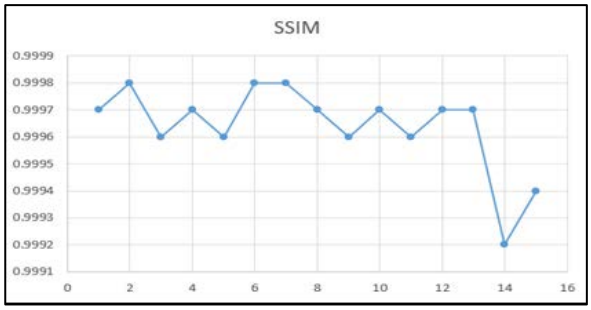


Fig. 11. SSIM Value.

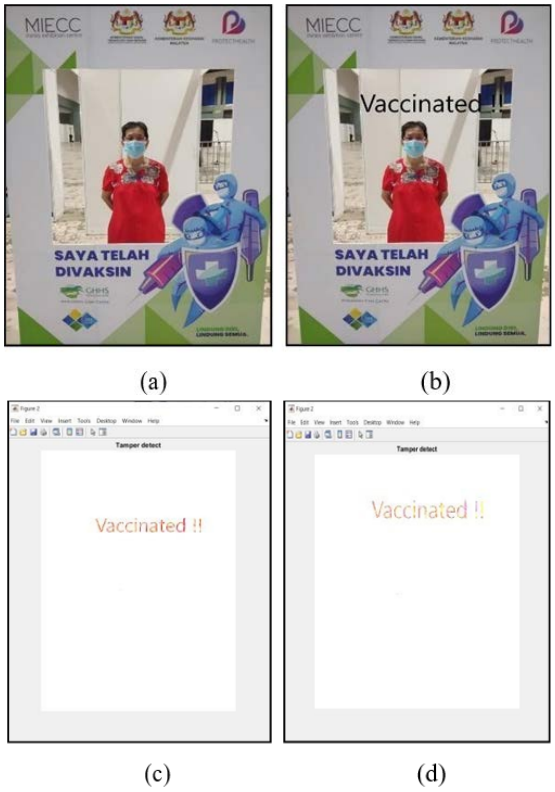


Fig. 12. (a) Original Image, (b) Tampered Image (c) Before Post, (d) After Posted.

In Fig. 13, a text "Sail boat" was inserted on the image (a) to produce a tampered image (b). After acquiring a tampered image, it was tested prior to upload in Instagram. Figure (c) is the result before uploading while figure (d) shows the result after uploading. Figure (a) was a general image taken from an image database, so the result shows no noise detected even after being uploaded to social media. The tamper was detected and marked it in red, as shown in Fig. 13(d).

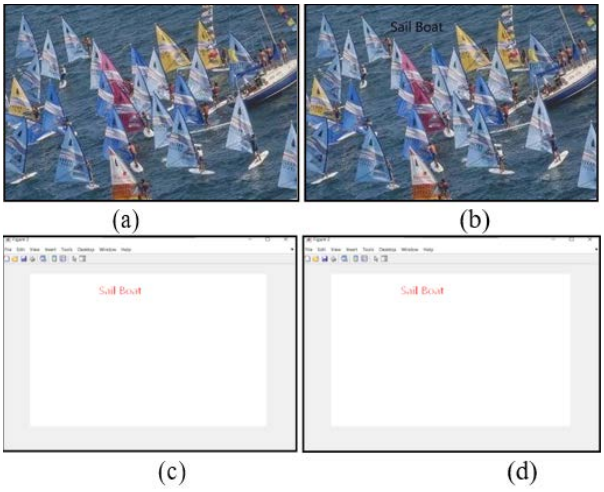
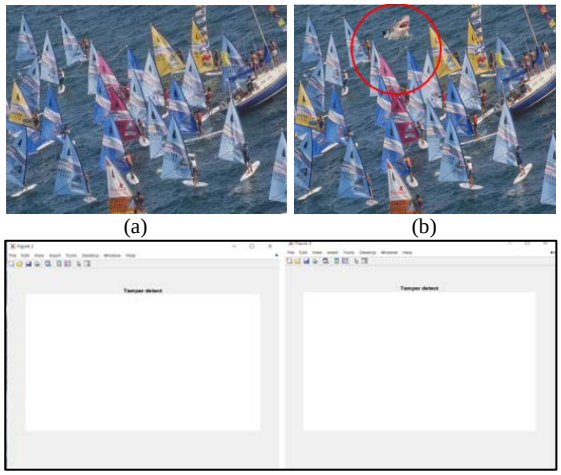


Fig. 13. (a) Original Image, (b) Tampered Image, (c) Before Post, (d) After Posted.

B. Image Splicing Attack

For image splicing attacks, Fig. 14(b) shows a spliced image. The objects were circled in red. Fig. 14(c) displays the results before uploaded into Instagram while Fig. 14(d) depicts the results after uploaded into Instagram. The results show that the tampered regions failed to be detected. JPEG compression applied by Instagram for uploaded images has not greatly affected the performance of the algorithm, since all the experiments provide identical result prior upload and after uploaded to Instagram platform.



(a) (b)

(c) (d)

Fig. 14. (a) Original Image, (b) Spliced Image, (c) Before Post, (d) After Posted.

Various attacks were done based on the studies revealing the common attacks to social media images. From the findings of the output, we can deduce that SPIRAL-LSB is able to detect text insertion attacks exclusively. Although some detection results display little fading after uploaded into social media. However, watermark is still intact with the host image as detection of tamper is successful. From the text insertion output, the deduction of the scheme was robust against JPEG compression because the tampered region is detected clearly after posted to Instagram.

From the 15 images that have image splicing attack, all of them were unable to be detected regardless of compression issue. Copy-move, cut-and-paste, and 3D-sticker insertion attacks produced identical outcomes. Hence, we can conclude that SPIRAL-LSB is inefficient to detect tamper for image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks. The watermarking scheme failed to detect the tampered region before the images posted to Instagram. Thus, SPIRAL-LSB is suitable for social media uploaded images that have been attacked by text insertion. For image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks, it is not suitable to be used.

The failure to detect for copy-move attack might be SPIRAL-LSB scheme was using comparison of intensity and parity bits to detect whether there was any tamper in the image or not. However, image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks may need the use of another approach to identify and key point-based forgery detection method have been proven helpful in detecting copy-move forgeries [31]. For image splicing, using two Markov features: coefficient-wise Markov features and block-wise Markov features in the discrete cosine transform (DCT) domain produce high detection accuracy [32]. Thus, SPIRAL-LSB was not effective to detect image splicing attack.

In this work, we used Least Significant Bit (LSB) which is a spatial domain technique. According to [33], the embedding of the watermark into the original image is done by selecting a subset of pixels and substituting the least significant bit of the selected pixels with the watermark bits. The LSB techniques, are easy to implement and requires a little computation cost for both embedding and extraction processes. On the hand, they are sensitive to signal processing operations and generally show reduced robustness to different attacks. Even though there are a large number of suggested LSB algorithms, there is still a lack of a robust solution, necessitating further study in this field.

Spatial domain techniques are simple and have a high payload, work directly on the pixel level, but these are not robust against various attacks [34]. In spatial domain the information is added simply by just varying the pixel values of the host signal. The values of some colors or pixels are also directly editable in the spatial domain techniques. In the least significant bit (LSB) substitution technique, the watermark is added in the least significant bit of each pixel. When the extraction of information is needed, the LSB of each pixel is read. However, the major disadvantage of this watermarking is that it is not robust again various attacks according to [35]. So, the weakness of least significant bit technique is shown in the experiments in this study. SPIRAL-LSB could not detect the tampered regions of image splicing, copy-move, cut-and-paste and 3D-sticker insertion attacks.

In our experiment, the performed attacks can be considered as pixel level tampering. Thus, SPIRAL-LSB algorithm is a block-wise technique, and it cannot detect pixel-level tampering. This drawback is called a localization problem and it was reported by [36] in 2002. Subsequently, fragile watermarking techniques have been developed to address localization problem [37, 38]. Recently, the authors [39] proposed two related fragile watermarking techniques. The first method is a statistical technique which is capable of detecting pixel-level tampering if the tampered area is small. The second one improves the tamper detection capability for a larger area by incorporating a hybrid of block-wise and pixelwise mechanism. However, the use of block information reduces its tamper resistance capability.

From the previous research done by other researchers, it is proven that LSB substitution techniques have weaknesses and limited robustness under various attacks such as lossy compression which implemented by Instagram sites. Instagram uses a lossy compression technique (JPEG compression) that reduces the image's quality and size to save storage space, reduce the amount of computing resources required for image processing, and speed up the loading or display of an image on a user's timeline. In comparison to the original image, the image posted on social media contains distortion and noise. Thus, the watermarking scheme also detected noise which is in yellow color in posted images.

As it worked in a spiral manner, which started at the center, the image processed should be in square size to ensure all the blocks were numbered. The scheme could only number the image blocks in the square which also led to generating the watermarking data in the square too, not in total if the image were in a rectangle shape. This limitation made the scheme not compatible with other social media sites images such as Facebook as Facebook support images vary in sizes.

V. Conclusion

Social media attacks represent the largest modern threat vector and are at all-high because roughly 3.5 billion people are on social media. Image splicing, copy-move, cut-and-paste, text, and 3D-sticker insertion were the most common types of attacks on social media. Social media platforms are often used for authentication to other website, applications, thus, this is a major attack vector. It can also be used to compromise various sectors for damage to reputation, operation, and financial gain. Hence, authentication on social media images is needed to protect the integrity of images.

This research has demonstrated that watermarking can provide authenticity for social media images. The fragile watermarking techniques for authentication with unique numbering, SPIRAL-LSB have been devised. This research has proven the existing techniques in fragile watermarking of color images by offering a way to embed in LSB in each plane of RGB without having the problem of less space or high data capacity. SPIRAL-LSB offers a novel way to number the blocks of the original image before being mapped while embedding. The spiral scan allows the data to be located farther and the operation time to be short. Although the watermarking scheme is only effective on text insertion attacks, it is proven to be robust against the effect of applying lossy compression, for instance JPEG, to such images. Despite the completion of this project, the necessity for more improvement in the future is required as the world is going through changes.

Acknowledgment

This research work is supported by a grant entitled ‘Authentication Watermarking in Digital Text Document Images using Unique Pattern Numbering and Mapping’ (RDU190366) and PGRS200369 supported by Universiti Malaysia Pahang.

References

[1] Barrett-Maitland, N., & Lynch, J. (2020). Social Media, ethics, and the privacy paradox. Security and Privacy from a Legal, Ethical, and Technical Perspective. https://doi.org/10.5772/intechopen.90906.

[2] Bullock, L., Gurd, J., & Hanlon, A. (2022, June 1). Global Social Media Statistics Research Summary 2022 [June 2022]. Smart Insights. Retrieved June 21, 2022, from https://www.smartinsights.com/social-media- marketing/social-media-strategy/new-global-social-media-research.

[3] Khan, M. N., Ashraf, M. A., Seinen, D., Khan, K. U., & Laar, R. A. (2021). Social Media for knowledge acquisition and dissemination: The impact of the COVID-19 pandemic on Collaborative Learning Driven Social Media Adoption. Frontiers in Psychology, 12. https://doi.org/10.3389/fpsyg.2021.648253.

[4] Lenhart, A. (2019, December 31). Chapter 4: social media and friendships. Pew Research Center: Internet, Science & Tech. Retrieved June 21, 2022, from https://www.pewresearch.org/internet/ 2015/08/06/chapter-4- social-media-and-friendships/.

[5] Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2017, November 6). Advances in social media research: Past, Present, and Future - Information Systems Frontiers. SpringerLink. Retrieved June 21, 2022, from https://link.springer.com/article/ 10.1007/s10796-017- 9810-y.

[6] Memon, A. M., Sharma, S. G., Mohite, S. S., & Jain, S. (2018, November). The role of online social networking on deliberate self-harm and suicidality in adolescents: A systematized review of literature. Indian journal of psychiatry. Retrieved June 21, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6278213/.

[7] Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2020, July 10). Setting the future of digital and social media marketing research: Perspectives and Research Propositions. International Journal of Information Management. Retrieved June 21, 2022, from https://www.sciencedirect.com/ science/article/pii/S026840122 0308082.

[8] Linas Jurkauskas. (2015, June). Digital piracy as an innovation in the recording industry - AAU. AALBORG UNIVERSITY. Retrieved February 8, 2022, from https://projekter.aau.dk/projekter/files/213765961/Digital_Pira cy_as_an_Innovation_in_Recording_ Industry_by_L._Jurkaus kas.pdf.

[9] Begum, M., & Uddin, M. S. (2020). Digital Image Watermarking Techniques: A Review.

[10] Cox, I. J., Miller, M. L. and Bloom, J. A. 2002. Digital watermarking. San Francisco. Morgan Kaufmann.

[11] Caldelli, R., Filippini, F. and Barni, M. 2006. Joint near- lossless compression and watermarking of still images for authentication and tamper localization. Signal Processing: Image Communication. 21:890–903.

[12] Chang, Chin-Chen, Yi-Hsuan Fan, and Wei-Liang Tai. 2008. Four-scanning attack on hierarchical digital watermarking method for image tamper detection and recovery. Pattern Recognition. 41(2): 654-661.

[13] Zain, J. M. and Fauzi, A. R. M. 2006. Medical Image Watermarking with Tamper Detection and Recovery. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3270-3273.

[14] Sahu, A. K., & Sahu, M. (2020). Digital Image Steganography and steganalysis: A journey of the past three decades. Open Computer Science, 10(1), 296–342. https://doi.org/10.1515/comp-2020-0136.

[15] Sreenivas, K., & Kamkshi Prasad, V. (2017). Fragile watermarking schemes for image authentication: A survey. International Journal of Machine Learning and Cybernetics, 9(7), 1193–1218. https://doi.org/10.1007/s13042-017-0641-4.

[16] Prasad, S., & Pal, A. K. (2020). Hamming code and logistic-map based pixel-level active forgery detection scheme using fragile watermarking. Multimedia Tools and Applications, 79(29-30), 20897–20928. https://doi.org/10.1007/s11042-020-08715-x.

[17] Rakhmawati, L., Suwadi, S., & Wirawan, W. (2020). Blind robust and self-embedding fragile image watermarking for image authentication and copyright protection with Recovery Capability. International Journal of Intelligent Engineering and Systems, 13(5), 197– 210. https://doi.org/10.22266/ijies2020.1031.18.

[18] Hurrah, N. N., Parah, S. A., Loan, N. A., Sheikh, J. A., Elhoseny, M., & Muhammad, K. (2019). Dual watermarking framework for privacy protection and content authentication of multimedia. Future Generation Computer Systems, 94, 654–673. https://doi.org/10.1016/j.future. 2018.12.036.

[19] Hisham, S. I., Muhammad, A. N., Badshah, G., Johari, N. H., & Mohamad Zain, J. (2016). Numbering with spiral patterns to prove authenticity and integrity in medical images. Pattern Analysis and Applications, 20(4), 1129–1144. https://doi.org/10.1007/s10044- 016-0552-0.

[20] Canales, K. (2021, February 9). Twitter surpassed 192 million daily active users in Q4 as the social media company grappled with criticism of its role in the spread of election misinformation. Business Insider.

Retrieved June 21, 2022, from https://www.businessinsider.com/twitter-earnings-q4- revenue-eps-new-users-2021-2.

[21] Schneider, G.M. & Gersting, J.L. 2004. Invitation to computer science. Course Technology.

[22] Lossy vs lossless compression - keycdn support. KeyCDN. (n.d.). Retrieved July 29, 2022, from https://www.keycdn.com/support/lossy-vs-lossless.

[23] Patvardhan, C., Kumar, P., & Vasantha Lakshmi, C. (2017). Effective color image watermarking scheme using ycbcr color space and QR code. Multimedia Tools and Applications, 77(10), 12655–12677.

[24] Prinkle Rani, & Jyoti Rani. (2015). Copy-move forgery attack detection in digital images. International Journal of Engineering Research And, V4(06). https://doi.org/10.17577/ijertv4is061110.

[25] Hassan, A., & Sharma, V. K. (2021). Texture based image splicing forgery recognition using a passive approach. International Journal of Integrated Engineering, 13(4). https://doi.org/10.30880/ijie.2021. 13.04.010.

[26] Aditi. (2020, September 16). Classical art memes: A visual analysis. openclosemag. Retrieved June 21, 2022, from https://www. openclosemag.com/post/classical-art-memes-a- visual analysis.

[27] Ross, A. S., & Rivers, D. J. (2019). Internet memes, media frames, and the conflicting logics of climate change discourse. Environmental Communication, 13(7), 975–994. https://doi.org/10.1080/17524032. 2018 .1560347.

[28] Gaaed, M., & Tahar, M. (2018). Digital Image Watermarking based on LSB techniques: A comparative study. International Journal of Computer Applications, 181(26), 30–36. https://doi.org/10.5120 /ijca2018918105.

[29] Zain, J. M. and Fauzi, A. R. M. 2006. Medical Image Watermarking with Tamper Detection and Recovery. 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3270-3273.

[30] Lin, C. and Chang, S. 2001. A Robust image authentication method distinguishing JPEG compression from malicious manipulation. IEEE Transactions on Circuits and Systems for Video Technology. 11(2): 153-168.

[31] Ulutas, G., &; Muzaffer, G. (2016). A new copy moves forgery detection method resistant to object removal with uniform background forgery. Mathematical Problems in Engineering,2016, 1–19. https://doi.org/10.1155/2016/321516.

[32] El-Alfy, E.-S. M., & Qureshi, M. A. (2014). Combining spatial and DCT based Markov features for enhanced blind detection of image splicing. Pattern Analysis and Applications, 18(3), 713–723. https://doi.org/10.1007/s10044-014-0396-4.

[33] Gaaed, M., & Tahar, M. (2018). Digital Image Watermarking based on LSB techniques: A comparative study. International Journal of Computer Applications, 181(26), 30–36. https://doi.org/10.5120/ijca 2018918105.

[34] Kumar, S., & Dutta, A. (2016). A novel spatial domain technique for digital image watermarking using block entropy. 2016 International Conference on Recent Trends in Information Technology (ICRTIT). https://doi.org/10.1109/icrtit.2016.7569530.

[35] Tao, H., Chongmin, L., Mohamad Zain, J., & Abdalla, A. N. (2014). Robust image watermarking theories and techniques: A Review. Journal of Applied Research and Technology, 12(1), 122–138. https://doi.org/10.1016/s1665-6423(14)71612-8.

[36] Fridrich, J., Goljan, M. and Du, R. (2001) Lossless data embedding – new paradigm in digital watermarking, EURASIP Journal of Applied Signal Processing, Vol.2, Pp. 185-196.

[37] Q. Li and N. Memon, “Security Models of Digital Watermarking," in Multimedia Content Analysis and Mining. vol. 4577, N. Sebe, Y. Liu, Y. Zhuang, and T. Huang, Eds., ed: Springer Berlin / Heidelberg, 2007, pp. 60-64.

[38] Shahreza, M. S. 2005. An Improved Method for Steganography on Mobile Phones.

[39] Zhang, H., Wang, C., & Zhou, X. (2017). Fragile Watermarking for image authentication using the characteristic of SVD. Algorithms, 10(1), 27. https://doi.org/10.3390/a1001002.