2021 9th International Symposium on Digital Forensics and Security (ISDFS) | 978-1-6654-4481-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISDFS52919.2021.9486339

A novel reversible fragile watermarking in DWT domain for tamper localization and digital image authentication

Gökhan Azizoğlu[1,2]
[1]*Department of Software Engineering*
Sivas Cumhuriyet University
Sivas, Turkey
gazizoglu@cumhuriyet.edu.tr, 4010930040@erciyes.edu.tr

Ahmet Nusret Toprak[2]
[2]*Department of Computer Engineering*
Erciyes University
Kayseri, Turkey
antoprak@erciyes.edu.tr

***Abstract*— In recent years, altering and tampering with digital images have become easier with the swift development of internet and computer technologies. Therefore, the use of image authentication in legal cases, digital forensic, and medical imaging has become of paramount importance. In this paper, as a solution to this problem, we propose an MD5 Hash-based blind reversible fragile watermarking method. We divide the input image into non-overlapped 4×4 blocks. Then, the watermark information generated from blocks using the MD5 hash function is embedded in the one-level DWT high and middle frequency sub-bands. Our experimental results demonstrate that the proposed method can provide reversibility for applications such as medical, forensic medicine, and military fields where the original image is essential. It can also provide the ability to resist various malicious attacks such as exchange of content, crop, text addition, copy-paste, rotation, content removal, and noise addition.**

Keywords—blind reversible fragile watermarking, tamper detection, image authentication, digital forensic

I. INTRODUCTION

Transmission and sharing of images have become very easy and fast with the development of communication and information technologies in recent years [1]. However, images can be easily tampered with by unauthorized people through image editing tools [2]. It is a critical issue to verify the integrity of sensitive images such as medical images, in which even the slightest distortion or alteration may cause misdiagnosis. It is also essential to detect the changes made to the image. Therefore, there is a need for a method that can verify the integrity of digital images and detect changes. Digital watermarking is one of the common methods used to solve this problem.

Digital watermarking techniques are classified into four major categories according to their robustness: robust, fragile, semi-fragile, and hybrid watermarking. Robust watermarking is a watermarking technique in which the watermark embedded in the image can resist attacks such as geometric transformation and image manipulation. It is used in copy control, copyright protection, fingerprinting, and broadcast monitoring applications. Fragile watermarking is a watermarking technique used in applications where the embedded watermark is broken at the slightest changes. Fragile watermarking methods are used to verify the content of the image and integrity. Semi-fragile watermarking techniques are fragile against malicious attacks while they are resistant to deliberate attacks on the embedded watermarking. Finally, hybrid watermarking methods include both fragile and robust watermarking. They are used in copyright protection, integrity verification, and image authentication [3-4].

Depending on the domain where the watermark is embedded, fragile watermarking can be split into two main categories: watermarking in the spatial domain and watermarking in the transform domain [5]. In the spatial domain watermarking methods, the watermark information is directly embedded in the pixel values of the cover image [6]. The most common spatial domain watermarking methods are the Least Significant Bit (LSB), Local Binary Pattern, and Histogram Modification [3]. In the transform domain watermarking, a transformation is first applied to the cover image to obtain transform coefficients. Then, the watermark data is embedded in obtained coefficients [6]. Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT) are the frequently used transformations in digital watermarking [5].

There are many fragile watermarking methods in the literature that aims at image authentication and tamper detection. These methods differ from each other in the number of blocks the images are divided into, being reversible and the domain they use.

Gül et al. proposed a fragile watermarking method that can detect alterations by dividing the input image into 16×16 blocks and recover alterations made up to 75%. This method uses the two least significant image bits for storing the watermark information [7].

Raj et al. proposed two fragile watermarking methods using the MD5 and SHA256 hash algorithms to detect alterations made to images. The first method divides the image into 8×8 blocks and then embeds the 128-bit MD5 hash value, which is used as watermark information, into the two LSB bits. In the

978-1-6654-4481-1/21/$31.00 ©2021 IEEE

second method, they embedded the SHA hash of 256 bits into the image using blocks of 16×16 [8].

Singh et al. proposed a DCT-based self-embedding fragile watermarking method for image authentication and recovery purposes. In the study, the image is divided into 2×2 blocks. Then, the 12-bit data obtained from these blocks is embedded using three LSBs of the corresponding pixels in the block map created [9].

Nguyen et al. proposed a reversible and highly accurate tamper detection method for image authentication purposes. In this study, the image is divided into 8×8 blocks, and the watermark data is embedded using the 2nd level DWT [10].

In this paper, a novel blind reversible fragile watermarking method based on the MD5 hash algorithm is presented. The proposed method divides the image into 4×4 blocks in order to high-precision tamper detection. Along with the increase in the amount of data to be stored in watermarking methods, the need for memory and associated costs increase. Bandwidth is a valuable asset, especially in network applications. The proposed method uses the MD5 hash algorithm to generate watermark data for each block. Then, it embeds the watermark data by modifying DWT coefficients. Thus, there is no need to store the original image and watermark. Accordingly, the amount to be transferred is reduced and the advantage of storage space is provided. Furthermore, the proposed method provides reversibility, which is crucial for application areas such as medical and military.

The main contributions of this work are as follows:

- Development of high-precision tamper detection and high watermarked image quality embedding technique.
- The proposed scheme supports blind extraction. Thus, as in other schemes, the memory space required to store the original image and watermark information is no longer needed.
- The reversibility of the proposed method makes it suitable for use in medical and military areas where the original image is crucial.

The rest of this paper is arranged as follows. In section II, the proposed method is discussed in detail. The experimental results and performance evaluations of the proposed method are presented in section III. Finally, the conclusion is given in section IV.

II. Proposed Method

We have designed an MD5 hash-based novel blind reversible watermarking method for tamper detection and digital image authentication. The proposed method includes watermark embedding and watermark extraction phases. The processes of watermark generation using the MD5 hash function and watermark embedding are discussed in section A. The extracting the embedded watermark, tamper detection, and tamper localization processes are discussed in section B.

A. Watermark Embedding Phase

In the algorithm of embedding the watermark which is shown in Fig. 1, the image is first divided into 4×4 blocks, and then all the block indexes are stored in a matrix called the block map. To ensure security, this block map, which will be used later when extracting the watermark, has been mixed with the Knuth shuffle algorithm, which is a modern key-based algorithm [11].

	Input: Original host image
	Output: Watermarked image
1:	Divide image into non-overlapped 4×4 blocks
2:	Blocks mixed with Knuth Shuffle algorithm
3:	Apply DWT to each block
4:	Calculate the average of the coefficients in the LL sub-band
5:	Obtain the binary hash value with a 128-bit length of the average value through MD5, and get the first 9 bits of this obtained hash value as watermark information (Wg)
6:	Embed the 9 bits of MD5 hash value to the HH, HL and LH subbands
7:	Obtain watermarked image through inverse DWT

Fig. 1. The watermark embedding algorithm

In the proposed method, the DWT coefficients of each block are modified to embed the watermark data. DWT is a widely used transformation in undetectable and robust watermarking. In DWT, each level of decomposition produces four bands of data defined as LL (low-low), LH (low-high), HL (high-low), and, HH (high-high). While a more robust watermarking is obtained when the watermark information is embedded in the LL sub-band, a higher imperceptibility can be achieved when embedded in the HH sub-band [12]. In the study, the HH sub-band and middle bands are used to embed the watermark, since it is aimed to achieve high imperceptibility.

In the proposed method, the watermark data is generated from the cover image, and the proposed method does not require the original image during watermark extraction. The transformed image is used to obtain watermark data without any processing to the LL sub-band of the blocks. To this end, the coefficients of the LL sub-band were averaged. Then, the binary hash value with a 128-bit length of the average value is obtained through MD5, and the first nine bits of this obtained hash value are used as watermark information (W_g).

To embed the obtaining watermark information into each block, one-level DWT first is applied. After applying DWT, four sub-bands (LL, LH, HH, and, HL) with coefficients in size 2×2 are obtained. In order to achieve minimum distortion in the image, the watermark is embedded using the absolute distance value calculated by using Eq. (1).

$$AD=|\ (a+b+c)/3\ - s|, \tag{1}$$

where s is the selected coefficient in a 2×2 block, a, b and c are adjacent coefficients. AD refers to the absolute distance. The coefficient matrix of 2×2 size is shown in Fig. 2.

s	a
b	c

Fig. 2. The coefficient matrix of 2x2 size

After the one-level DWT is applied, three bits are embedded in each of the obtained HH, LH, and HL sub-bands with a total of nine bits. The first digit of the calculated absolute distance after the comma is used to embed watermark information. In order to increase the capacity in each sub-band, three bits of data are embedded instead of one bit. To achieve this, three-bit data is first expressed in a decimal number. Then, the numbers after the comma are shifted one unit to the right and the obtained number is placed as in the first digit after the comma. In this process, against the problem of the number being a zero neutral element, the numbers that can occur are relocated to the range of one to eight by increasing the value of the number by one.

After embedding the watermark information in the absolute distance, Eq. (2) is used to obtain the new coefficient.

$$Y = \begin{cases} \bar{x}\text{-}AD', \text{If } \bar{x} > 0 \text{ and } s \geq 0 \text{ and } \bar{x} > s \\ \bar{x} + AD', \text{If } \bar{x} \geq 0 \text{ and } s \geq 0 \text{ and } \bar{x} \leq s \\ \bar{x} + AD', \text{If } \bar{x} < 0 \text{ and } s \geq 0 \\ \bar{x}\text{-}AD', \text{If } \bar{x} < 0 \text{ and } s < 0 \text{ and } \bar{x} > s \\ \bar{x} + AD', \text{If } \bar{x} < 0 \text{ and } s < 0 \text{ and } \bar{x} \leq s \\ \bar{x}\text{-}AD', \text{If } \bar{x} \geq 0 \text{ and } s < 0 \end{cases}, \quad (2)$$

where $\bar{x}$ is the average of adjacent coefficients (a, b, and c), *AD'* is the new absolute distance value in which the information of the watermark is embedded, and *Y* is the new coefficient value obtained.

The obtained new coefficient is replaced with the selected coefficient (s) in the 2×2 block and applied inverse DWT transformation to reach watermarked block. After all blocks are watermarked, the blocks are merged to produce a watermarked image. The general block diagram for embedding the watermark information into the image is shown in Fig. 3.

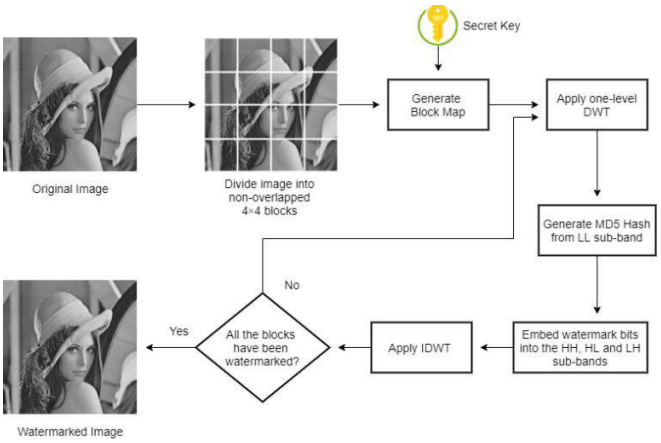

Fig. 3. Block diagram of the proposed watermark embedding process

B. Watermark Extracting Phase

Fig. 4 gives the watermark extraction algorithm. In the extraction phase, similar to the watermarking process, the watermarked image is first divided into 4×4 blocks. Then, the block map is generated with the help of the secret key used in the embedding of the watermark. 2×2 sized HH, LH, LL, and HL sub-bands are obtained by applying one-level DWT to each block according to the block map. The first coefficient of the sub-band is selected, and the absolute distance is calculated using Eq. (1). The first digit of the absolute distance contains the embedded data after the comma is taken. The taken number is reduced by one and converted to binary to reach the embedded watermark bits. The same procedures are applied for all the HH, HL, and LH sub-bands. At last, the watermark data is obtained by combining extracted bits. After the watermark data is received, absolute distance is restored to the state before the watermark was embedded; thus, the original image is obtained again.

Input: Watermarked image
Output: Extracted image and tampered locations
1: Divide image into non-overlapped 4×4 blocks
2: Mix the blocks with the Knuth Shuffle algorithm using the secret key
3: Apply DWT to each block
4: Calculate the average of the coefficients in the LL subband
5: Calculate the MD5 hash value of coefficients average (W_g)
6: Extract the data embedded in HH, HL and LH sub-bands (W_e)
7: Compare W_g and W_e, depending on the comparison create 4×4 blocks as tampered or non-tampered.
8: Obtain extracted image block and tampered block through inverse DWT
9: After extracting all the watermarks, merge the image and tampered blocks that will indicate the tampered locations

Fig. 4. Watermark extracting algorithm

The watermark data embedded in the block is obtained by using the LL sub-band coefficients (W_g) to detect the tampering. If the embedded watermark data of a block is not equal to extracted watermark data, it is then marked as tampered. Fig. 5 shows the general steps of the watermark extraction process.

III. Experimental Results

In this section, imperceptibility analysis and tamper detection under different known malicious attacks experiments were conducted to demonstrate the performance of the proposed method. A set of standard grayscale 512×512 size Baboon, Barbara, Boat, Lena, and Peppers images were used.

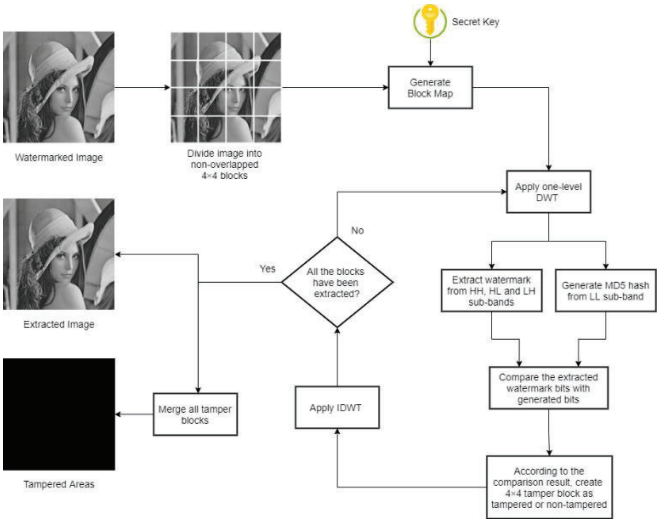

Fig. 5. Block diagram of the proposed watermark extraction process

A. *Imperceptibility Analysis*

Peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and mean square error (MSE) image quality metrics were used to measure the perceptual quality of the watermarked and restored images. In all of the following equations, I, I_w, and $H \times W$ represent the original image, the watermarked image, and image dimension, respectively.

Mean square error measures the mean square difference between pixel values of the two images. The closer the two images are to each other, the lower the mean square error. The MSE is measured by Eq. 3 [3].

$$MSE(I, I_w) = \frac{1}{HW} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} (I(i,j) - I_w(i,j))^2 \quad (3)$$

PSNR is used to measure the similarity between the original image and the watermarked image. The higher PSNR value reveals that the amount of similarity between the two images has increased [3].

$$PSNR(I, I_w) = 10 * \log 10 \frac{max_I^2}{MSE} \quad (4)$$

where: max_I represents the maximum intensity value of the input image. Since 8-bit gray-level images are used in this study, max_I is 255.

SSIM is used to measure the structural similarity of images. The SSIM value varies between -1 and 1. An SSIM value of 1 indicates that the watermarked image is exactly similar to the original image [3].

$$SSIM(I, I_w) = \frac{(2\mu_I \mu_{I_w} + c_1)(2cov + c_2)}{(\mu_I^2 + \mu_{I_w}^2 + c_1)(\sigma_I^2 + \sigma_{I_w}^2 + c_2)} \quad (5)$$

where: μ_I and μ_{I_w} are the average of I and I_w, σ_I^2 and $\sigma_{I_w}^2$ are the variances of I and I_w, respectively. cov is the covariance of I and I_w. c_1 and c_2 are balancing constants [3].

TABLE I. PERCEPTUAL QUALITY OF WATERMARKED IMAGES

Image	*PSNR(dB)*	*MSE*	*SSIM*
Baboon	64.5589	0.0228	0.9999
Barbara	64.5264	0.0229	0.9999
Boat	64.5399	0.0229	0.9999
Lena	64.5099	0.0230	0.9998
Peppers	64.5499	0.0228	0.9998
Average	64.537	0.02288	0.99986

The perceptual quality of watermarked images is shown in Table I. Perceptual qualities of watermarked images with the proposed method references average values were 64.537, 0.02288, and 0.99986, respectively, for PSNR, MSE, and SSIM.

The histogram comparison for the image obtained after the watermark is extracted from the watermarked Lena image is shown in Fig. 6. The histogram results illustrate that after the watermark is extracted, the original image can be obtained without distortion.

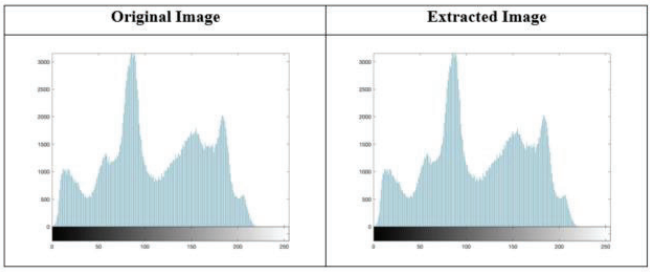

Fig. 6. Histogram of original and extracted image

A comparison of the proposed method with other image authentication schemes using Lena image is given in Table II. The compared schemes Raj et al.[8], Gül et al. [13], Singh et al. [14], Huang et al. [15] and Lee et al. [16] are not reversible and they have lower perceptual quality than the proposed method. Lo and Hu [17] used the same number of blocks as this scheme and proposed a reversible method, but they achieved lower perceptual quality.

TABLE II. COMPRASION OF PROPOSED SCHEME WITH OTHER IMAGE AUTHANTICATION SCHEMES

Schemes	*Block Size*	*PSNR (dB) of Watermarked Image*	*Reversibility*
Huang et al. [15]	4×4	33.724	No
Gül et al. [13]	2×4, 4×2	38.3545	No
Raj et al. [8]	8×8	44.16	No
Singh et al. [14]	1×2	46.82	No
Le et al. [16]	4×4	47.20	No
Lo and Hu [17]	4×4	51.454	Yes
Proposed	4×4	64.537	Yes

B. *Performance Against Tampering Attacks*

The performance of the proposed method against tampering attacks was evaluated by applying exchange of content, crop, copy-paste, text addition, rotation, content removal, and noise addition attacks to watermarked images.

Exchange of content attack is the replacement of a portion taken over a watermarked image with another portion taken over the same or another watermarked image. It is seen in Fig. 7 that the proposed method, Gül et al.'s [13] method, and Singh's [14] method can detect the tamper made, but Lee's [16] method cannot. In addition, it seems that the proposed method can detect tampers with higher accuracy compared to other schemes.

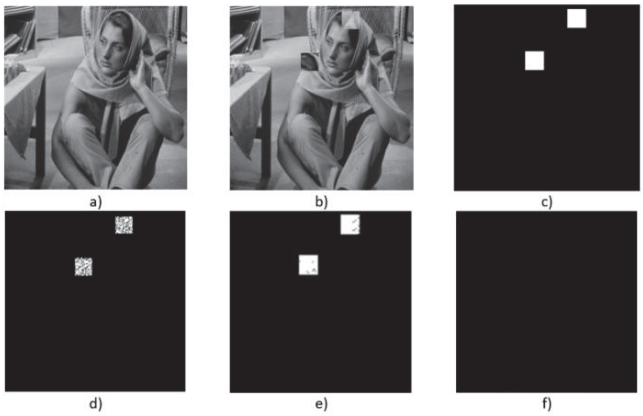

Fig. 7. Exchange of content attack: a) Watermerked Barbara, b) Attacked image, c) Proposed method, d) Gül et al.'s [13] method, e) Singh et al.'s [14] method, f) Lee et al.'s [16] method

Comparison of the proposed scheme with other related schemes resistance against %15 cropping attack is shown Fig. 8.

Fig. 8. %15 cropping attack: a) Watermerked Lena, b) Attacked image, c) Proposed method, d) Gül et al.'s [13] method, e) Lee et al.'s [16] method, f) Singh et al.'s [14] method

Some known malicious attacks, such as copy-paste, content removal, text addition, rotation, and noise addition have been applied to the watermarked image, and the results are shown in Fig. 9. The results show that the proposed method can resist these attacks and detect changes with high accuracy.

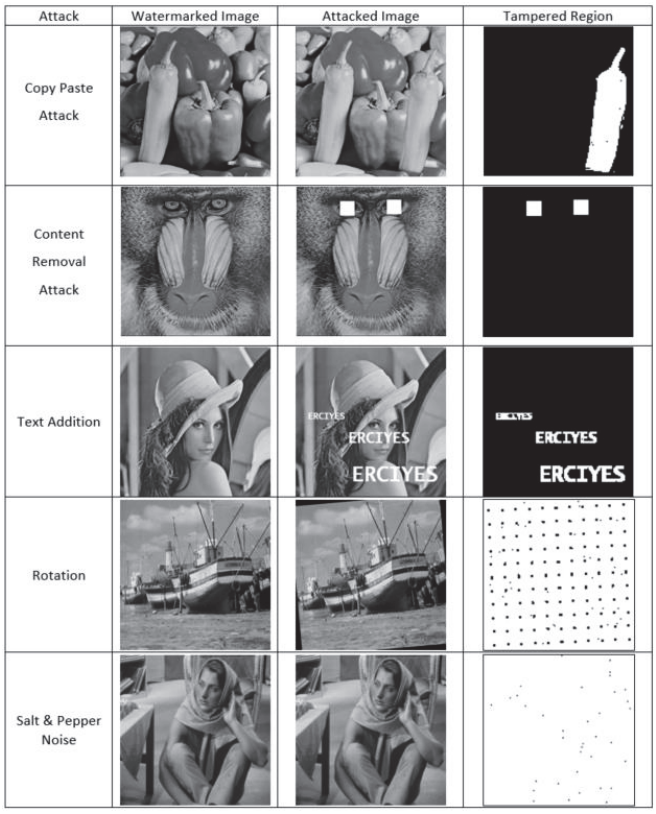

Fig. 9. Performance against some known malicious attacks

The proposed method, stored the watermarked image in the double data type, as in transform-based methods in the literature [10]. Therefore, the watermarked image is of double type its use in some fields such as DICOM image format in medical applications is prevented.

IV. Conclusion

This paper proposes a novel blind reversible fragile watermarking method, which provides a high-quality watermarked image and detects tampering with high accuracy. Input image is divided into 4×4 non-overlapped blocks and watermark generated by using MD5 hash function. Then, the generated watermark embedded in the DWT domain. In the proposed method average values were obtained 64.537, 0.02288, and 0.99986, respectively, for PSNR, MSE, and SSIM. Tamper detection is determined by comparing the embedded watermark and the hash value produced from the watermarked image.

Experimental results illustrate that the proposed method can detect tampering with high sensitivity in the presence of attacks such as exchange of content, content removal, crop, copy-paste, text addition, rotation, and noise addition. Besides, it does not need the original image during watermark extraction. Also, generating the watermark from the original image provides an advantage in terms of storage space.

REFERENCES

[1] P. Selvam, S. Balachandran, S. Pitchai Iyer, and R. Jayabal, "Hybrid transform based reversible watermarking technique for medical images in telemedicine applications," Optik, vol. 145, pp. 655–671, Sep. 2017.

[2] Y. C. Hu, C. C. Lo, and W. L. Chen, "Probability-based reversible image authentication scheme for image demosaicking," Futur. Gener. Comput. Syst., vol. 62, pp. 92–103, Sep. 2016.

[3] A. F. Qasim, F. Meziane, and R. Aspin, "Digital watermarking: Applicability for developing trust in medical imaging workflows state of the art review," Computer Science Review, vol. 27. Elsevier Ireland Ltd, pp. 45–60, Feb. 01, 2018.

[4] S. M. Mousavi, A. Naghsh, and S. A. R. Abu-Bakar, "Watermarking Techniques used in Medical Images: a Survey," Journal of Digital Imaging, vol. 27, no. 6. Springer New York LLC, pp. 714–729, Nov. 06, 2014.

[5] E. Gul and S. Ozturk, "A novel hash function based fragile watermarking method for image integrity," Multimed. Tools Appl., vol. 78, no. 13, pp. 17701–17718, Jul. 2018.

[6] A. Anand and A. K. Singh, "Watermarking techniques for medical data authentication: a survey," Multimed. Tools Appl., pp. 1–33, Apr. 2020.

[7] E. Gul and S. Ozturk, "A novel triple recovery information embedding approach for self-embedded digital image watermarking," Multimed. Tools Appl., pp. 1–26, Aug. 2020.

[8] N. R. N. Raj and R. Shreelekshmi, "Blockwise Fragile Watermarking Schemes for Tamper Localization in Digital Images," in 2018 International CET Conference on Control, Communication, and Computing, IC4 2018, Nov. 2018, pp. 441–446.

[9] D. Singh and S. K. Singh, "DCT based efficient fragile watermarking scheme for image authentication and restoration," Multimed. Tools Appl., vol. 76, no. 1, pp. 953–977, Jan. 2017.

[10] T. S. Nguyen, C. C. Chang, and X. Q. Yang, "A reversible image authentication scheme based on fragile watermarking in discrete wavelet transform domain," AEU - Int. J. Electron. Commun., vol. 70, no. 8, pp. 1055–1061, Aug. 2016.

[11] Knuth, Donald E. (1969). Seminumerical algorithms. The Art of Computer Programming. 2. Reading, MA: Addison–Wesley. pp. 139–140.

[12] K Swaraja, K Meenakshi, and P. Kora, "An optimized blind dual medical image watermarking framework for tamper localization and content authentication in secured telemedicine," Biomed. Signal Process. Control, vol. 55, p. 101665, Jan. 2019.

[13] E. Gul and S. Ozturk, "A novel pixel-wise authentication-based self-embedding fragile watermarking method," Multimed. Syst., vol. 1, p. 3, Feb. 2021.

[14] P. Singh and S. Agarwal, "An efficient fragile watermarking scheme with multilevel tamper detection and recovery based on dynamic domain selection," Multimed. Tools Appl., vol. 75, no. 14, pp. 8165–8194, Jul. 2015.

[15] C. C. Lin, Y. Huang, and W. L. Tai, "A novel hybrid image authentication scheme based on absolute moment block truncation coding," Multimed. Tools Appl., vol. 76, no. 1, pp. 463–488, Jan. 2017.

[16] C. F. Lee, J. J. Shen, Z. R. Chen, and S. Agrawal, "Self-embedding authenticationwatermarking with effective tampered location detection and high-quality image recovery," Sensors, vol. 19, no. 10, May 2019.

[17] C. C. Lo and Y. C. Hu, "A novel reversible image authentication scheme for digital images," Signal Processing, vol. 98, pp. 174–185, May 2014.