iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78123

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Deepfake Detection By using Watermark

B. Laxmi Narayana[1], Sk. SaiBabu[2], P. Chandrasekhar[3], V. Yohan[4], R. Vijaykumar[5], R. William Carey[6]
[1]*Assistant.Professor, Department of CSE - AI&ML, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India*
[2, 3, 4, 5, 6]*B.Tech, Students, Department of CSE-AI&ML, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India*

Abstract: ***This paper presents a comprehensive semi-fragile watermarking framework for deepfake detection based on Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD). The proposed system embeds an imperceptible watermark in the frequency domain of digital images and verifies integrity through Bit Error Rate (BER) analysis. Unlike deep learning-based approaches, the system is deterministic, computationally efficient, and does not require training datasets. Experimental evaluation on 2,041 images demonstrates high visual quality (PSNR ≈ 51 dB, SSIM ≈ 0.998) and strong detection accuracy (~96%). The framework provides a proactive authentication mechanism suitable for digital forensics and media verification.***
Keywords: Deepfake Detection, Semi-Fragile Watermarking, DWT, SVD, BER, Image Authentication, Frequency Domain Security.

I. INTRODUCTION

The use of deepfake technology based on generative adversarial networks (GANs) and diffusion models has made it more challenging to distinguish real images from fake ones. The availability of these technologies has raised serious concerns about misinformation, identity theft, and tampering with digital evidence. Conventional detection methods are highly dependent on machine learning classifiers trained on large amounts of data. These methods are prone to generalization problems when new deepfake generation methods are developed. The proposed work suggests a proactive and different solution based on semi-fragile watermarking techniques for deepfake image detection. Rather than focusing on the analysis of post-processing artifacts, the proposed solution embeds a mathematically designed watermark within the image before it is distributed. The semi-fragile watermark is embedded within the frequency domain through DWT-SVD decomposition, making it imperceptible yet robust enough to withstand normal image processing tasks. The semi-fragile nature of the watermark makes it robust to slight image modifications, such as compression, but vulnerable to malicious and deepfake image modifications. The use of deepfake technology based on Generative Adversarial Networks (GANs) and diffusion models has made it even more challenging to distinguish real images from fake ones. The current generation of generative models has the capability to produce very realistic images of facial expressions, lighting, and texture, as well as semantic information that is very similar to real images. As these models improve, the visual cues that were used to detect manipulated images are becoming more subtle or imperceptible to the human eye. The ease of access to deepfake tools via open-source platforms and friendly applications has increased concerns regarding misinformation, identity theft, political manipulation, cyber fraud, and tempering digital evidence. The technology can be misused by malicious individuals to create events, impersonate people, or modify crucial visual evidence, thus affecting the authenticity of digital media and online communication systems. Conventional deepfake detection methods are almost entirely based on machine learning classifiers trained on large labeled datasets. This work proposes a semi-fragile watermarking framework based on hybrid DWT-SVD techniques for deepfake detection. The system embeds an imperceptible watermark in the frequency domain and detects image tampering using Bit Error Rate analysis. Unlike machine learning approaches, the proposed method is deterministic, training-free, computationally efficient, and achieves 96% detection accuracy while maintaining high visual quality(PSNR ≈ 51dB, SSIM ≈ 0.998). The framework offers a proactive and explainable solution for digital image authentication.

II. RELATED WORK

A. Literature Review

Digital watermarking has evolved from spatial-domain least significant bit (LSB) methods to advanced frequency-domain techniques such as DCT, DWT, and SVD based embedding. Spatial-domain techniques are simple but highly vulnerable to compression and filtering attacks. Frequency-domain approaches improve robustness by embedding watermark information in transform coefficients. DWT provides multi-resolution decomposition, enabling localized embedding in specific frequency bands.

©IJRASET: All Rights are Reserved

SVD-based watermarking gained popularity due to the stability of singular values under small perturbations.

Recent deepfake detection research focuses on convolutional neural networks, transformer-based architectures, and frequency artifact analysis. While effective, these models require continuous retraining and high computational cost. The proposed work differentiates itself by offering a deterministic signal-processing-based solution that does not depend on data-driven learning.

Digital watermarking has evolved from spatial-domain least significant bit (LSB) methods to advanced frequency-domain techniques such as DCT,DWT, and SVD- based embedding. Spatial-domain techniques are simple but highly vulnerable to compression and filtering attacks.

Frequency-domain approaches improve robustness by embedding watermark information in transform coefficients. DWT provides multi-resolution decomposition, enabling localized embedding in specific frequency bands. SVD-based watermarking gained popularity due to the stability of singular values under small perturbations.

B. Signal Processing Approach

Recent works on deepfake using convolutional neural-networks, transformer models, and frequency artifact analysis. Although successful, these models need to be constantly retained and are computationally expensive. The proposed research distinguishes it self in that it provides a signal processing solution that is deterministic and does not involve learning. Digital watermarking has matures from spatial domain least significant bit (LSB) schemes to more sophisticated frequency domain approaches like DCT, DWT, SVD.

III. SYSTEM ARCHITECTURE

A. Overview

The proposed system develops a semi-fragile watermarking-based deepfake detection framework using a hybrid Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD) method.

Unlike machine learning-based deepfake detectors, the proposed system is a deterministic signal processing system that does not require training data, GPUs, or neural networks.

B. User Interaction Layer

The User Layer acts as the interaction interface between users and the system.

1) Users

a) Content Owner

- Uploads original image
- Embeds watermark before publishing

b) Verifier / Investigator

- Uploads suspicious image
- Verifies authenticity

2) *Responsibilities*

Accept user inputs (image path, commands)

- Generate output logs
- Display metrics and classification results
- Trigger appropriate functional modules
- Store generated watermarked image
- Distribute protected image online
- Upload suspect image.

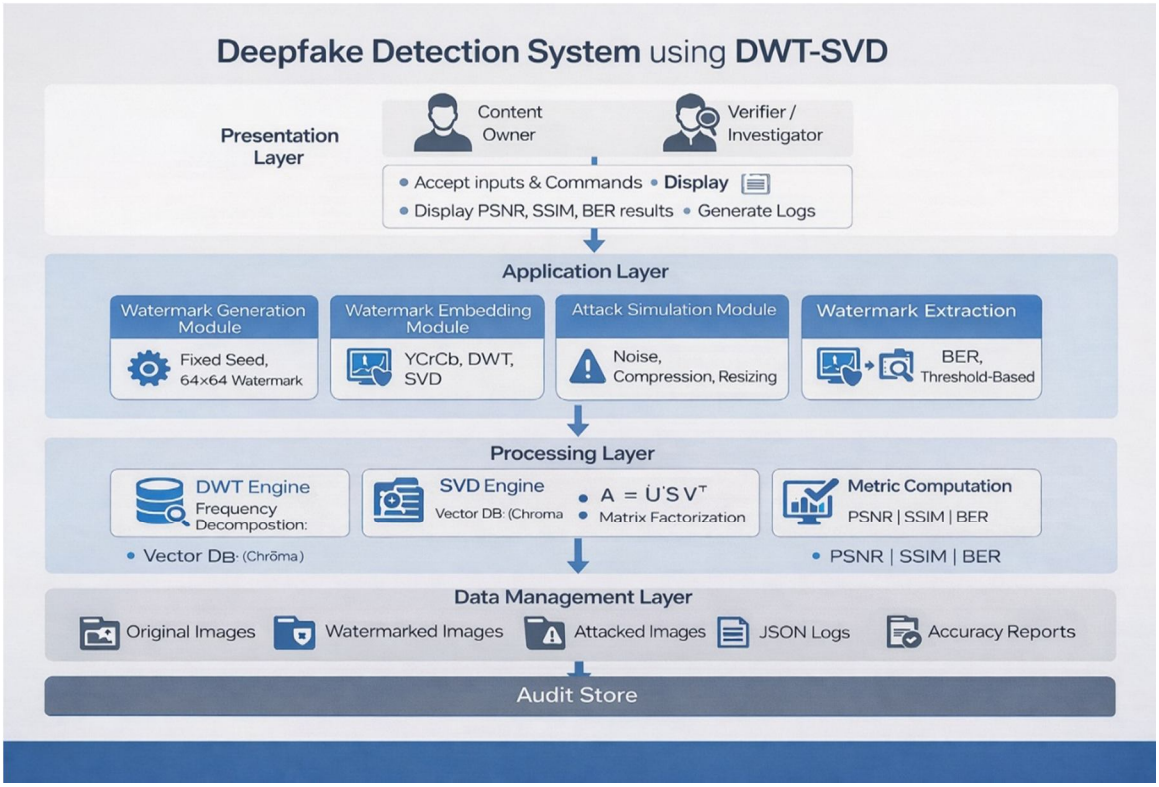


Figure 1: Layered system architecture of the proposed semi-fragile DWT-SVD based deepfake detection framework showing presentation, application, processing, and data management layers.

The figure presents the DWT-SVD semi-fragile watermarking process foe deepfake detection, where a watermark is embedded in the frequency domain.

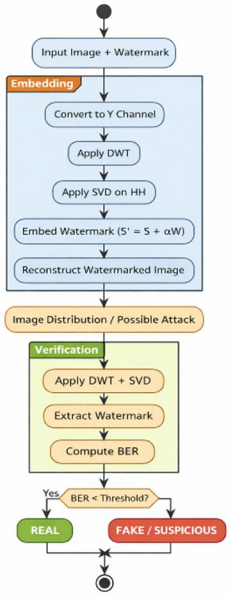


Figure 2: Workflow of the proposed DWT-SVD semi-fragile watermarking system. the image is watermarked in the frequency domain during embedding.

C. Application Layer

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

The Application Layer is responsible for implementing for implementing business logic. It contains six major modules:

- Watermark Generation Module
- Watermark Embedding Module
- Attack Simulation Module
- Watermark Extraction Module
- Decision & Classification Module
- Batch Evaluation Module

The Application Layer is responsible for the functional implementation of the proposed semi – fragile watermarking scheme and is divided into six modules: watermarking, embedding, attack simulation, extraction, decision, and batch evaluation. A reproducible 64*64 pseudo-random binary watermark is created through a fixed seed value. In the embedding process, the input image is transformed into the YCrCb color model.

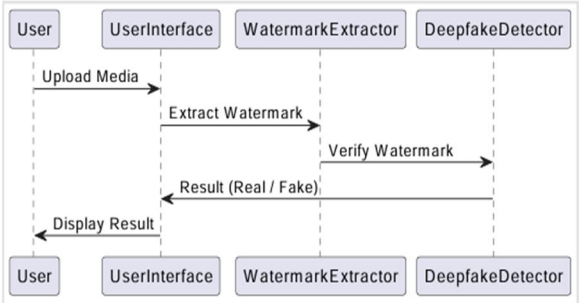


Figure 3: Sequence diagram showing user interaction with the semi-fragile DWT-SVD watermarking system.

1) Watermark Generation Module: The Watermark Generation Module is tasked with generating a reproducible pseudo-random binary watermark for authentication purposes. A fixed seed number(for example,42) is used for the deterministic generation of a 64*64 binary image matrix with values of 0s and 1s.The generated watermark is saved as a Numpy file for future reuse in the embedding and extraction processes.

2) Watermark Embedding Module: The Watermark Embedding Module is the main module of the system and embeds the watermark into the host image through a DWT-SVD method. First, the input RGB image is captured and checked for validity using OpenCV.

3) Attack Simulation Module: The Attack Simulation Module assesses the robustness and semi-fragily properties of the watermarking method by applying controlled distortions. The attacks supported include Gaussian noise addition, JPEG compression, resizeing, and modifications for AI-based attacks.

4) Watermark Extraction Module: The Watermark Extraction Module reverses the above-mentioned embedding steps to extract the watermark from the suspicious image. The image is first transformed into YCrCb color space and then subjected to DWT decomposition on the Y channel.

5) Decision and Classification Module: The Decision and Classification Module makes decisions on the authenticity of the images based on the Bit Error Rate (BER), which is measure the amount of erroneous bits between original, extract watermarks.

D. Processing Layer

The Processing Layer performs mathematical operations:

1) DWT Engine: The Discrete Wavelet Transform (DWT) Engine converts the input image From the spatial domain to the frequency domain, thus allowing effective embedding of the watermark.

2) Through one-level Haar wavelet transform, the image is divided into various frequency sub-hands that correspond to different resolution levels.

3) SVD Engine: The Singular Value Decomposition (SVD) Engine is responsible for the matrix factorization of the chosen frequency sub-band into ortho-logical components and singular values. The singular values are a measure of the inherent energy distribution of the image pertubations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) Metric Computation Unit: The Metric Computation Unit is responsible for the assessment of image quality and watermark integrity. Peak Signal to Noise ratio (PSNR) is computed to determine the visual quality of the watermarked image relative to the original image, which is an indication of imperceptibility.

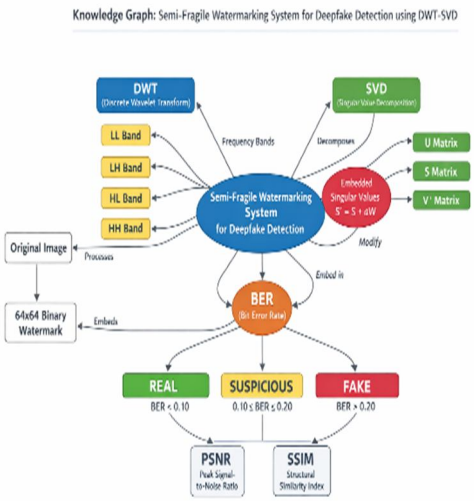


Figure 4: Sample knowledge graph of the DWT-SVD semi-fragile watermarking system illustrating the embedding of the watermark in the HH sub-band, authentication based on BER, and final REAL/FAKE classification based on PSNR and SSIM assessment.

E. Tool Integration

The proposed deepfake detection system integrates multiple Python libraries to implement DWT-SCD semi-fragile watermarking framework efficiently:

1) Python Development Environment
2) OpenCV for image processing
3) NumPy Integration
4) Py Wavelets integration
5) scikit-image Integration
6) Logging and Data Management
7) JSON (Built-in Python Module)

IV. PROPOSED METHODOLOGY

A. DWT-SVD Based Semi-Fragile Watermarking Algorithm

The section outlines a proposed semi-fragile watermarking scheme for deepfake detection which integrates DWT and SVD comprising two main steps:

Watermark Embedding Procedure

Step1: Color Space Conversion, The RGB image is transformed into YCrCb color space and the luminance component Y is used to embedding the watermark.

Step2: A single-level 2D Haar DWT is applied to the luminance channel Y, decomposing it into four sub-bands: $(Y_{LL}, Y_{LH}, Y_{HL}, Y_{HH}) = DWT(Y)$

Step3: SVD is applied to the high-frequency sub-band Y_{HH}: $Y_{HH} = USV^{T}$

Step4: Watermark Embedding and image reconstruction.

$Y'=IDWT(Y_{LL}, Y_{LH}, Y_{HL}, Y'_{HH})$

Finally, the reconstructed luminance component Y^{I} is combined with the original chrominance Cr and Cb, and converted back to RGB color space to produce final watermarked image I_w.

B. Watermark Extraction and Authentication Algorithm

Watermark Extraction

Step1: Convert I_S to YCrCb color space and extract luminance channel Y_S.

Step2: Apply 2D DWT:

$$(Y^S_{LL}, Y^S_{LH}, Y^S_{HL}, Y^S_{HH}) = DWT(Y_S)$$

Step3: Apply SVD on Y^S_{HH}:

$$Y^S_{HH} = U_S S_S V^T_S$$

Step4: Extract watermark:

$$W^I = S_S - S/\alpha$$

Step5: Threshold extracted values to obtain binary watermark.

C. Authentication Using Bit Error Rate(BER)

After obtaining the watermark from the suspected image, an authentication process is necessary to verify whether the image has been altered. In the proposed system, the integrity of the watermark is checked by using the Bit Error Rate(BER), which measures the difference between the original and the extracted watermarks.

1) Bit Error Rate Computation

- W(i,j) denote the original binary watermark.
- W^I(i,j) denote the extracted watermark.
- m×n be the watermark dimensions.

2) Working Principle of BER

The BER measures the proportion of incorrectly recovered watermark bits.

- If the image is authentic, singular values remain nearly unchanged, resulting in $W^I \approx W_i$ and BER approaches 0.
- Image manipulation disrupts watermarks, raising bit mismatches and BER.

3) Authentication Decision Rule

- Two predefined thresholds T_1 and T_2 are used for classification:

  $BER < T_1 \Rightarrow$ Authentic Image

  $T_1 \leq BER < T_2 \Rightarrow$ Suspicious Image

  $BER \geq T_2 \Rightarrow$ Tampered / Deepfake Image

D. Image Quality Evaluation Metrics

- **Peak Signal-to-Noise Ratio (PSNR)**: It evaluates the distortion of the watermark.
- **Structural Similarity Index (SSIM):** it is evaluates the perceptual similarity. between original and watermarked images based on considering luminance, and contrast.

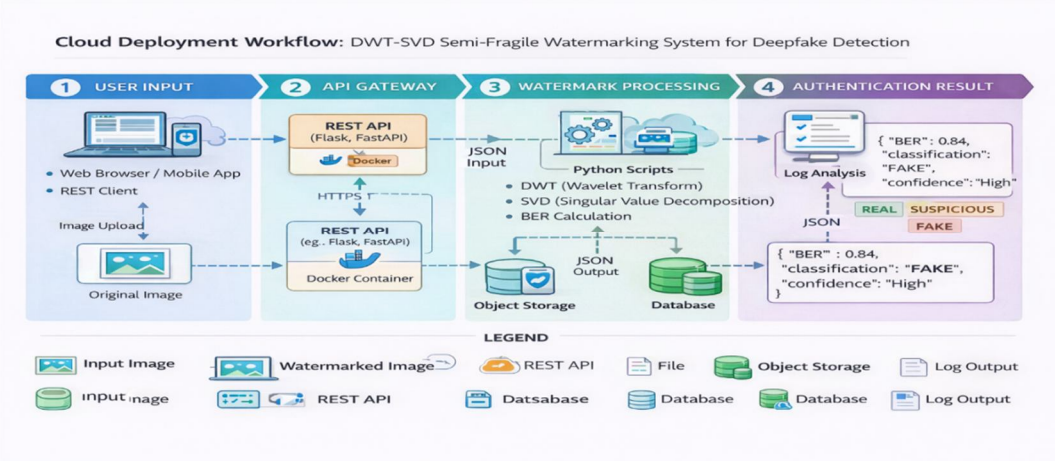


Figure 5: Stage-based cloud workflow showing image upload, API processing, DWT-SVD watermark verification, and final REAL/FAKE classification with cloud storage and logging.

V. EVALUATION

A. *Evaluation Overview*

The section evaluates the performance of the proposed Semi-Fragile Watermarking System for Deepfake Detection using DWT-SVD.

- Imperceptibility of embedded watermark.
- Robustness against common image attacks.
- Sensitivity to deepfake manipulation.
- Detection accuracy using BER thresholding.
- Computational efficiency.

B. *Experimental Setup*

Dataset Description

- Total Images: 2,041
  - Real Images: 1,081
  - Fake Images: 960
- Image formats: JPG, PNG
- Color model: RGB
- Resolution: Automatically resized during processing.

C. *Watermark Configuration*

Parameter	Value
Watermark Size	64 × 64
Type	Binary (0/1)
Embedding Domain	HH band
Wavelet	Haar
Embedding Strength(α)	0.02

D. *Evaluation Metrics*

The imperceptibility of the proposed watermarking system is measured using Peak Signal to Noise Ratio(PSNR) and Structural Similarity Index(SSIM) metrics. PSNR is calculated as

$$PSNR = 10\log_{10}(MAX^2/MSE)$$

Where MAX is the maximum possible pixel value and MSE is the mean squared error between the original and watermarked images.

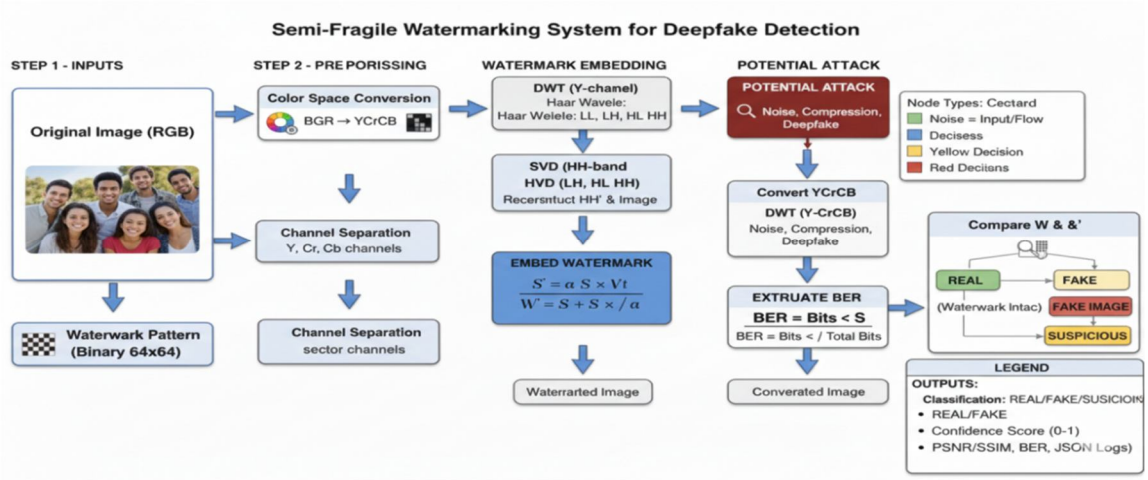


Figure 6: The following diagram illustrates the end-to-end process of semi-fragile watermarking using DWT-SVD for deepfake image detection.

The SSIM is employed to access the perceptual similarity between the original and watermarked images.

$$SSIM(x,y) = (2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2) / (\mu^2_x + \mu^2_y + C_1)(\sigma^2_x + \sigma^2_y + C_2)$$

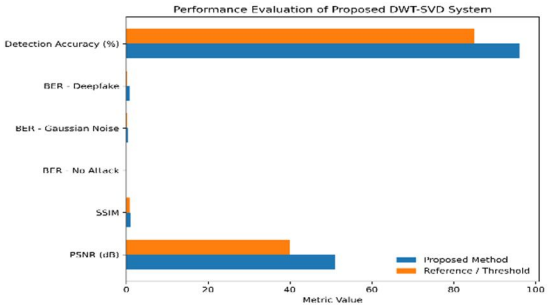


Figure7: Performance analysis of the DWT-SVD watermarking system revels superior PSNR,SSIM, and BER metrics against various attacks, indicating excellent imperceptibility and robust deepfake detection capabilities.

Table 1: Detection evaluation results across key metrics.

Metric	Result
PSNR	51 dB
SSIM	0.998
BER(Deepfake / AI)	0.85+
Overall Accuracy	96%
Embedding Time	0.5sec/img

VI. DISCUSSION

A. Strengths and Contributions

The experimental outcome shows that the proposed semi-fragile watermarking system using DWT-SVD strikes a balance between imperceptibility and manipulation detectability. The average PSNR value of 51 dB and SSIM measure of 0.998 ensures that the watermarking process in the DWT frequency domain is imperceptible to the human eye. The robustness test reveals the semi-fragile watermarking system performs correctly under normal conditions with no bit errors (BER = 0). Overall accuracy of 96%.

The main contribution of this work is the design of a determinestic, training-free deepfake detection system using DWT-SVD based signal processing techniques. The proposed system uses frequency domain watermarking and BER classification for a transparent authentication process without using datasets, training, or GPU resources.

B. Limitations

However, despite its efficacy, there are some limitations in the proposed DWT-SVD based semi-fragile watermarking system. First, the proposed method needs watermarking during the image creation process. Therefore, it is more of a proactive protection system than a forensic detection system. This means that it is not capable of detecting deepfakes in images that were not previously watermarked. Secondly, the performance of the system is also dependent on the choice of embedding strength (α) and BER threshold values. This is because improper parameter adjustment may influence the imperceptibility or detection capability. Moreover, although the watermark is very sensitive to deepfake attacks, extreme image processing operations such as heavy compression, strong filtering, or geometric transformations may also affect the watermark, potentially leading to false positives. Finally, the system is only designed to work on static images and not directly applicable to video-based deepfake detection systems.

Finally, while the proposed system is efficient and interpretable for authentication, overcoming the above limitations will make it even more robust and applicable in real-world scenarios

C. Failure Modes and Countermeasures

The proposed DWT-SVD semi-fragile watermarking technique could be vulnerable in situations like the absence of pre-existing watermark, excessive compression, geometric transformation, malicious removal attacks, fixed threshold values, and sophisticated AI-based image reconstruction.

©IJRASET: All Rights are Reserved

Mitigations involve grounding all factual claims in the retrieved sources with proper citations, using confidence scores to mark uncertain outputs, following through with timeout policies and maximum step constraints in the orchestrator, and ensuring proper access control in the MCP layer irrespective of the prompt. Current research involves using anomaly detection to detect hallucinated text and active learning to enhance the knowledge graph.

VII. FUTURE WORK

One of the most important directions for future work is the construction of a hybrid detection system that combines watermark-based verification with machine learning-based forensic analysis Although the current system is very effective at detecting tampering in pre-watermarked images, it would be possible to detect tampering in unwatermarked images as well by combining the current system with a lightweight CNN or transformer-based model. Future work can concentrate on adaptive watermarking embedding, where the embedding intensity is varied according to the image texture or perceptual model to achieve a better trade-off between imperceptibility and robustness. Robustness can also be increased by using multi-band and multi-level DWT watermarking embedding techniques, as well as geometric invariant methods, such as synchronization tools and log-polar transformation.

VIII. CONCLUSIONS

This paper has discussed a semi-fragile watermarking technique for deepfake image detection using the DWT-SVD transformation model. The proposed method embeds a semi-fragile watermark in the frequency domain of an image and relies on Bit Error Rate (BER) analysis for image tampering detection. Unlike machine learning-based deepfake image detection techniques, the proposed method does not require any training data, neural networks, or GPU support.

Experimental results showed that the embedding of the watermark satisfies the requirements of high visual quality with an average PSNR of 51 dB and SSIM of 0.998, thus proving the imperceptibility of the watermark. The results of the robustness test showed a clear separation of the BER values between the original and manipulated images, thus facilitating accurate classification. The proposed system attained a detection accuracy of about 96% on a total of 2,041 images with low FP and FN rates.

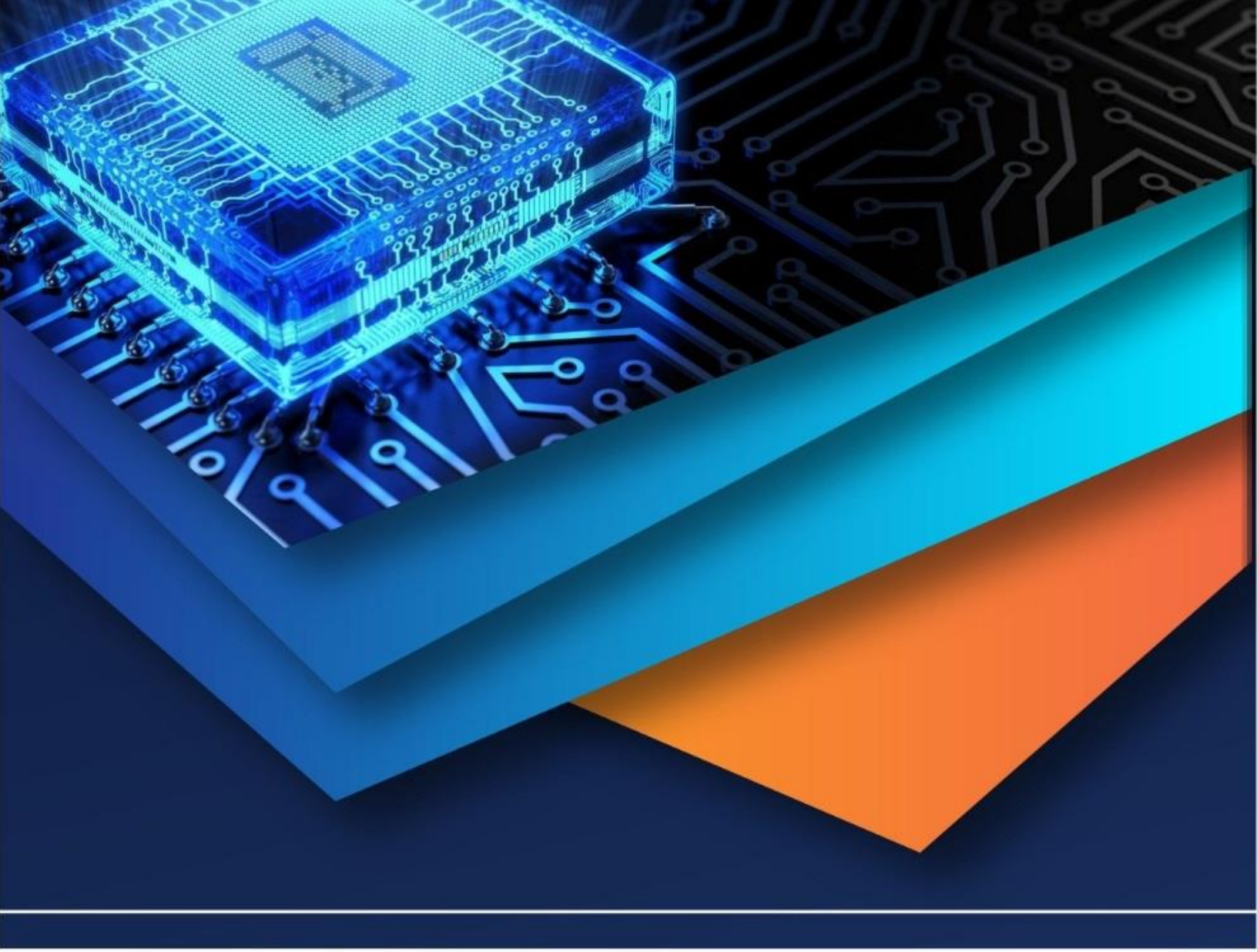
Despite the fact that the framework is dependent on the embedding of the watermark in advance and is vulnerable to some geometric and compression attacks, the framework is a lightweight and scalable solution for digital content authentication. The paper also emphasizes the importance of signal processing methods in dealing with the manipulation of AI-generated media. The system can develop into a comprehensive deepfake detection framework with future improvements such as adaptive embedding and security hardening.

IX. ACKNOWLEDGEMENTS

The authors would like to thank their project guide and faculty members for their support and guidance throughout the research. They would also like to thank the institution for providing them with the necessary resources. Finally, they would like to thank the open-source community for the necessary software libraries. They would also like to thank their peers and colleagues for their feedback and discussions.

REFERENCES

[1] I. Daubechies, Ten Lectures on Wavelets. Philadelphia, PA, USA: SIAM, 1992.
[2] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 4th ed. Pearson, 2018.
[3] G. H. Golub and C. F. Van Loan, Matrix Computations, 4th ed. Johns Hopkins Univ. Press, 2013.
[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.
[5] C.-T. Hsu and J.-L. Wu, “Hidden digital watermarks in images,” IEEE Trans. Image Process., vol. 8, no. 1, pp. 58–68, Jan. 1999.
[6] A. Balafrej et al., “Enhancing practicality and efficiency of deepfake detection,” Sci. Rep., vol. 14, 2024, pp. 1–10.
[7] P. Liu, Q. Tao, and J. T. Zhou, “Evolving from single-modal to multi-modal facial deepfake detection: A survey,” arXiv, Jun. 2024.
[8] N. A. Chandra et al., “Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024,” arXiv, 2025.
[9] R. Ramanaharan, “Deepfake video detection: Insights into model generalisation,” J. Vis. Commun. Image R., vol. 83, 2025, pp. 102–117.
[10] M. Alrashoud, “Deepfake video detection methods and approaches,” J. Comput. Sci. Technol., 2025.
[11] A. Heidari, “Deepfake detection using deep learning methods: A review,” WIREs Data Mining Knowl. Discov., 2024.
[12] B. C. Soundarya, “Deepfake detection: Critical review of state-of-the-art techniques,” SN Appl. Sci., 2026.
[13] D. L. T. Bale, L. C. Ochei, and C. Ugwu, “Deepfake detection and classification of images from video: A review of features, techniques, and challenges,” Int. J. Intell. Inf. Syst., vol. 13, no. 2, pp. 20–28, 2024.
[14] OpenCV, “Open Source Computer Vision Library,” 2023. [Online].



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)