



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Explainable Hybrid Gan–Rednet Framework For Bit-Level Medical Image Steganalysis

Solanki Pattanayak^{1*}, Payal Bose², Apratim Mitra³

¹Research Scholar, Department of Computer Science, Swami Vivekananda University, Barrackpore, India, Email: solankiphd2023@gmail.com, 0009-0003-0990-0785

²Assistant Professor, Department of Computer Science and Engineering, Swami Vivekananda University, Barrackpore, India, Email: bosepaya1@gmail.com, 0000-0002-2177-5775

³Professor, Department of Computer Applications, Haldia Institute of Technology, Haldia, India, Email: apratim777@yahoo.com, 0000-0002-7785-9978

* Corresponding Author: solankiphd2023@gmail.com

ABSTRACT

With the increasing digitization of medical imaging in healthcare, protecting sensitive patient data has become critical. Steganography, the process of hiding information within images, poses significant security risks, necessitating advanced steganalysis techniques for medical images. Traditional methods rely on handcrafted features and shallow learning, limiting their performance on high-resolution and bit-level embedded data. This study proposes an Explainable Hybrid GAN–REDNet Framework for bit-level medical image steganalysis and security enhancement. The framework integrates generative adversarial networks (GAN) for encoding-decoding and data augmentation with Residual Enhanced Discriminative Networks (RED-Net) for precise stego detection, ensuring both high fidelity and robust classification. A multimodal brain tumor dataset (CT and MRI) was used, comprising 5,000 images. Preprocessing included normalization, augmentation, and generation of stego images via multi-bit LSB embedding. GAN-based augmentation enriched feature diversity, while RED-Net discriminated cover and stego images at the pixel and bit levels. Bit-plane sensitivity analysis identified lower bit-planes (0–2) as critical for detection. Explainable AI techniques, including GradCAM, were employed to interpret model focus areas. The hybrid framework achieved 86.11% accuracy, 0.86 F1-score, and ROC-AUC of 0.9989, outperforming baseline REDNet (78.42% accuracy). Confusion matrix analysis confirmed high stego detection sensitivity, and GradCAM visualizations aligned with bit-plane sensitivity findings, validating interpretability. The proposed model enhances medical image security by combining generative and discriminative paradigms, enabling robust, interpretable, and bit-level steganalysis, critical for ethical and secure healthcare imaging.

KEYWORDS: Medical image steganalysis, GAN, RED-Net, Bit-plane analysis, Stego detection, Explainable AI (XAI), Data security.

1. Introduction

In the digital age of healthcare and telemedicine, medical imaging has been inseparable in the clinical diagnosis, disease monitoring, and treatment planning [1]. The growth of sensitive medical information transfer and communication has grown exponentially due to the growing use of digital imaging technologies like Magnetic Resonance Imaging (MRI) [2], Computed Tomography (CT) [3], and Ultrasound. As the number of cloud-based storage tools and Internet of Medical Things (IoMT) devices are increasingly used [4], the prevention of unauthorized access, tampering, and information leakage of medical images has become a significant issue. Amongst the most “elusive and possibly dangerous dangers to the integrity of images is the steganographic ones, which are the techniques of hiding information in a digital image in a manner inaccessible to human perception [5]. The matching countermeasure is called medical image steganalysis that aims at identifying and examining the concealed information within medical images to guarantee the confidentiality, integrity, and reliability of healthcare information systems [6].

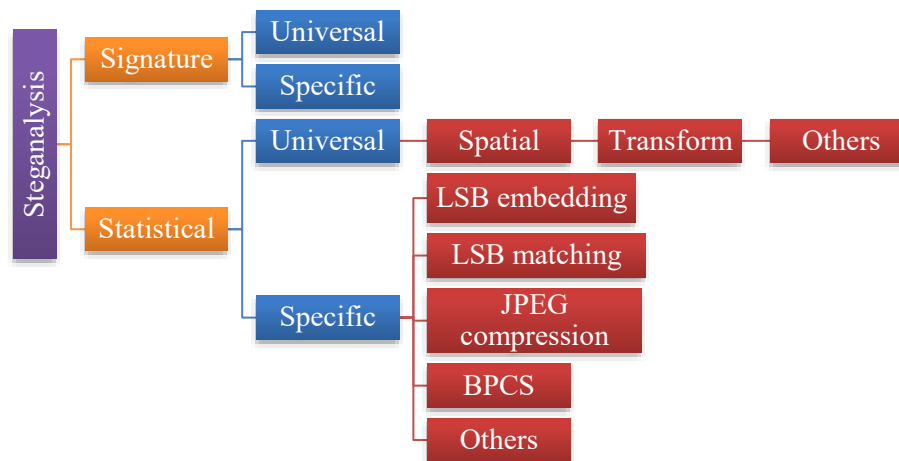


Figure 1: Classification of image steganalysis methodologies.

Figure 1 provides a clear and structured classification of image steganalysis methodologies, visually demonstrating the progression from traditional handcrafted statistical-feature-based techniques to modern machine- and deep-learning approaches. This classification helps readers quickly understand how the field has evolved, the relative strengths and limitations of each method, and why traditional approaches—despite their earlier success—struggle with complex, high-resolution medical images or advanced embedding schemes. By presenting all methodological categories together, the figure supports the discussion on current challenges and highlights existing research gaps, thereby framing the need for more robust, adaptive, and data-driven steganalysis models.

Traditional methods used for steganalysis generally rely heavily on manually curated statistical features as well as shallow learning algorithms to identify unusual image patterns. While these approaches have yielded relatively good performance, they are typically restricted in their ability to assess complex, high-resolution medical data, or applied with complex embedding methods to the pixel or bit level [7]. The emergence of machine learning (ML) and deep learning (DL) has changed this picture, as they present models that are data-driven and can automatically derive hidden representations, and complicated spatial-spectral features automatic to medical imaging [8, 9]. In particular, deep neural networks are specifically useful for identifying small anomalies and other irregularities that occur as a product of embedding operations [10]. Because medical images vary widely and steganographic distortions are extremely subtle, models must be scalable and robust to accurately detect pixel-level changes [11]. The steganalysis methods through deep learning are categorically classified into the generative and discriminative-based ones. Generative models such as Generative Adversarial Networks (GANs) rely on finding the data underlying distribution, and adversarial learning based on generating image patterns [12, 13]. By simulating realistic image content, these models can synthesize and recognize the structures normally changed during data embedding. This makes the GANs highly suitable in research of invisible manipulations of pixel brightness and spatial co-occurrence injected by steganographic methods. On the other hand, models like Convolutional Neural Networks (CNNs) [14], Support Vector Machines (SVMs), and Residual Enhanced Discriminative Networks (RED-Net) are designed to classify images based on learned decision boundaries delimiting cover and stego samples. The models perform extremely well to discriminate between patterns and identify anomalies particularly when changes are statistically prominent but not visually discriminable [15].

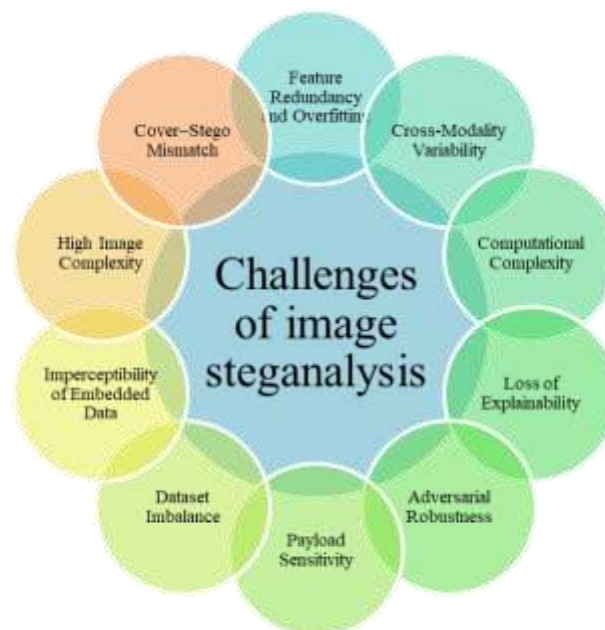


Figure 2: Key Challenges of IS (Image Steganalysis)

Regardless of individual merits, both generative and discriminative models have their own limitations to some extent, individually. Generative networks are capable of well-imitation of visualization features yet can fail to recognize intricate embedding properties whilst the discriminative networks are well-informed of classification boundaries yet tend to lack the capacity to convert or introspect structural beginnings of the hidden data [16]. In order to address these difficulties, scientists are turning towards hybrid systems which incorporate generative and discriminative learning systems into a single system. The models provide both detection and reconstruction at the same time where the system has the capability to provide synthesis-based knowledge and classification-based accuracy [17]. The bit-level interpretation of medical images is extremely important in this hybrid framework.

Every pixel in a digital image is an ordered series of binary bits, and steganographic algorithms typically exploit this representation by hiding information in the LSBs in a way that results in the least possible perceptual distortion [18, 19]. Although does result in no visible changes to the medical images, the pixel-level modifications result in perceptible change to the reliability of the imaging diagnostic, and potentially to confidentiality of the patient data. To begin to detect such fine-grained changes, models that can work at a very low noise rate while being highly sensitive to small statistical differences are needed [20]. The combination of a generative model such as GAN for encoding-decoding patterns and a discriminative model such as RED-Net for fine-grained anomaly detection forms an efficient basis for such precision-based work.

Therefore, this study aims to critically examine existing medical image steganalysis techniques, with a particular focus on bit-level detection challenges in clinical imaging. The study seeks to analyze the strengths and limitations of traditional, machine-learning, and deep-learning approaches, and to evaluate the emerging role of hybrid generative-discriminative frameworks such as GAN-REDNet models in improving detection accuracy and robustness. Additionally, it highlight the importance of bit-plane analysis, explore the contribution of explainable AI methods in enhancing model transparency, and identify key research gaps that need to be addressed to ensure secure, reliable, and ethically responsible medical image handling. The objectives of this study are -

- To develop and validate an explainable hybrid GAN-REDNet framework for robust bit-level steganalysis in medical CT/MRI images, outperforming baseline REDNet by achieving 86.11% accuracy.
- To conduct comprehensive ablation studies quantifying the synergistic contributions of GAN augmentation and REDNet discrimination across component variants.
- To evaluate detection performance across extended payloads (0.1-4.0 bpp) and real stego algorithms (LSB, PVD, Wow, JUniwad) with systematic SOTA benchmarks (e.g., vs. SRNet, XuNet).
- To perform hyperparameter optimization using Bayesian methods, establishing optimal configurations for reproducible convergence.
- To analyze bit-plane sensitivity and integrate GradCAM-based XAI for interpretable model focus aligned with embedding artifacts.
- To demonstrate clinical utility through multimodal generalization, security-prioritized bias (high stego recall), and robustness metrics for IoMT/telemedicine deployment.

2. Review Of Literature

Over the past decade, research relating to steganalysis (the detection of data hidden in images) in medical image analysis has progressed to deep learning-based approaches that outperform traditional statistical based solutions for detection. Beginning studies on security in medical images via watermarking by Liao et al. (2018) [21] and Elbasi et al. (2018) [22] used frequency domains or watermarking-based steganalysis systems. Elbasi et al. (2018) [22] maintained inter-block relationships in JPEG medical images that surpassed traditional DCT-based models, whilst Elbasi et al. utilized a dual-domain watermarking model that was resilient to traditional attacks such as compression and noise. Mansour and Abdelrahim (2019) [23] proposed a ripplet transform based steganalysis system to provide use in telemedicine, to improve resilience to RS attacks, leaving behind both purely statistic-based models and even to offer transform-based, steganalysis-based solutions. In their study, Khalifa et al. (2019) [24] implemented what they called co-occurrence matrix features and frequency analysis features with backpropagation neural networks, achieving a detection rate demonstrating 81.82 detection rates ratios at a 0.5 bpp hiding ratio. Collectively, all studies reported on the value of frequency and transform domains provided some resilience while showing restrictions on the scalability, or amounts of embedding patterns or complexity of embedding patterns.

Along with the development of the sphere, discriminative deep learning models quickly turned into the prevalent schemes applied in medical image steganalysis. Indicatively, to identify pixel-level changes, the authors of the studies in the field preceding this study, including the study by Ozcan et al. (2020) [25], utilized simple discriminative models herself by coming up with the concept of SVM-based steganalysis that provided up to 99.28 percent accuracy with LSB-based hidden data. Conversely, such three-bit image steganography algorithm as that proposed by Siddiqui et al. [26] for medical and e-health care systems can be demonstrated to have a more sophisticated and more complex implementation, assuming the availability of better disguise imperceptibility, the lower the accuracy of any discernment model. As part of that, Hemalatha et al. (2021) [27] suggested a variant of deep CNN that was built on the AlexNet and implemented a pre-processing method that featured histogram equalization and Shearlet transformed data to enhance classification results. Correspondingly, Karakis (2021) [28] and more recently, Karakis et al. (2023) [29], extended this to ensemble deep learning models that still provided robustness yet were still trying to achieve accuracy and generalization, with the use of both ResNet and DenseMaps, even with the conditions of cover-stego mismatch. Although these discriminative models demonstrate accuracy and capability of generalization, this is accompanied by a lack of explainability and data imbalance which will restrict the use of caseload and application into a clinical setting.

In parallel, generative models began to reshape the landscape of medical image steganography and steganalysis. Yang et al. (2018) [30] was the first to train adversarial networks on U-Net architectures, demonstrating that generative networks would learn more effective ways to embed and recover complex reconstruction processes compared to the traditional ones. Adding to this, Zhang et al. (2019) [31] proposed an approach to an invisible steganography using GANs where performance is characterized by a high payload capacity and minimal perceptual quality, and Yuan et al. (2022) [32] proposed to use adversarial perturbations as part of GAN models to respond to steganalyzers based on CNNs. It is also the same with the Ambika and Uplaonkar (2024) [33] who created an attention-directed GAN to transform medical images, where diagnostic features were not altered during data encoding. Complementarily, GANs were used together with transfer learning in Al-Obaidi et al. (2024) [34] to compute 85 percent classification on unbalanced stego and non-stego datasets, which is a direct response to one of the critical issues in generative steganalysis. These papers all highlighted the strengths of GANs to generate invisible yet securely embed data, but also revealed their susceptibility to adversarial examples and overfitting to small datasets.

In the recent past, hybrid models integrating generative and discriminative paradigms have been sought in an effort to achieve an optimal balance between synthesis and detection precision. Chen et al. (2023) [35] initially introduced the RED-Net (Residual and Enhanced Discriminative Network), which minimized steganalysis mistakes in the spatial and apparent JPEG domains. Parallel to this, Balaji et al. (2024) [36] presented RGAN-Net, the residual generative-discriminative network for high-texture medical images which resulted in a significant decrease in steganalysis errors. Chen et al. (2025) [37] later advocated a fuzzy deep learning-based generative model to hide patient information as unnoticeable noise within a revealed image without compromising segmentation accuracy. Together, these hybrid works, overall, mark a departure from merely considering isolated models as being discriminative or generative towards the convergence of generative and discriminative paradigms with a focus on both robust detection and understandable reconstruction. The scientific community increasingly comes to appreciate that in medical image security, the optimal outcomes arise when generative synthesis and discriminative analysis are combined and coordinated with each other for better performance and understandability.

3. Research Methodology

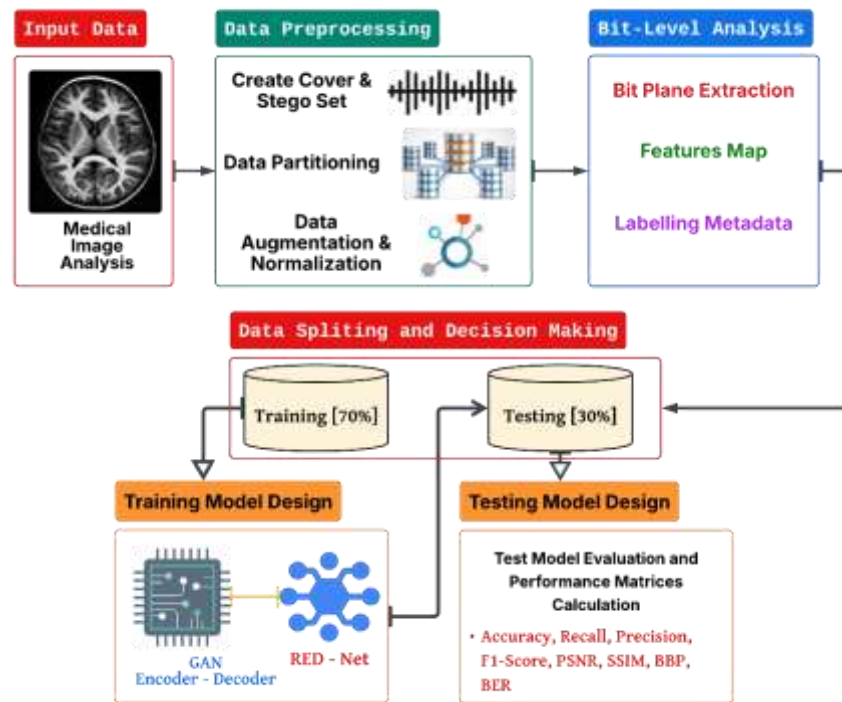


Figure 3: Proposed model

3.1 Dataset Collection

In this study, the Brain Tumor Multimodal Image (CT and MRI) Dataset acquired from Kaggle [38] is used as the principal reference for medical images. This dataset consists of a highly comprehensive and varied collection of Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans that are useful for implementing and assessing the proposed Explainable Hybrid GAN-REDNet Framework. Incorporating both imaging types ensures that the model can learn and migrate over diverse imaging features, which bolsters bit-level steganalysis. Each image in the data set has been all annotated with the imaging modality and whether or not it has a brain tumor, which makes is suitable for supervised learning and testing of both generative and discriminative processes. High-quality images allow for effective and rigorous pixel-level and bit-level analysis, critical in examining minimal changes made during the embedding process.

Table 1: Overview and Sample Instances of the Brain Tumor Multimodal Image (CT and MRI) Dataset

Attribute	Description / Example
Dataset Name	Brain Tumor Multimodal Image (CT and MRI)
Source	Kaggle
Image Modalities	CT Scans, MRI Scans
Number of Images	Approximately 5,000 total (2,500 MRI and 2,500 CT images)
Image Resolution	Ranges from 256×256 to 512×512 pixels
Classes	Tumor Present, Tumor Absent
Sample Instance 1 (MRI)	<i>MRI_001.jpg</i> → Modality: MRI → Label: Tumor Present
Sample Instance 2 (CT)	<i>CT_045.png</i> → Modality: CT → Label: Tumor Absent
Applications	Multimodal learning, tumor classification, image synthesis, bit-level embedding, and steganalysis
Usage in Study	Used to train and evaluate GAN for data embedding/reconstruction and RED-Net for discriminative steganalysis across CT and MRI modalities

3.2 Data Pre-processing

The preprocessing step is an integral part of the proposed Explainable Hybrid GAN-REDNet Framework to ensure that input medical images are normalized and ready for both the generative and discriminative learning stages. The data used is medical images like MRI, CT, and ultrasound scan images, which are considered as the baseline or cover images. Then, corresponding stego images are formed by embedding secret data at the bit level based on Least Significant Bit (LSB) and multi-bit LSB embedding. For every pixel intensity $I(x,y)$, the stego pixel $I'(x,y)$ is calculated as:

$$I'(x,y) = I(x,y) - (I(x,y) \bmod 2^n) + b_i \dots \dots \dots (1)$$

where n indicates the number of modified bits, which can range from 1–3 bits, and b_i refers to the secret data bit. The altered bit position and payload capacity (0.1–1.0 bpp) will help to generate more stego versions of the image

to represent different embedding scenarios, while the paired cover and stego images will be labeled to efficiently train a discriminative RED-Net model.

By applying controlled data augmentation, we can increase the variability of the training data without compromising the model's robustness and generalization capability. Controlled augmentations will only be applied to the training data set. Augmentations will include small rotations ($\pm 5^\circ$), horizontal and vertical flips, and a small intensity scale, or changes in intensity, while any anatomical or diagnostic structures should not change mathematically. Mathematically, an affine transformation is:

$$I_a(x', y') = I(A \cdot [x, y]^T + t) \dots\dots\dots(2)$$

where A is the transformation matrix and t is the translation vector. This enhances the variability of training samples without altering the underlying clinical content.

Finally, normalization is performed to standardize image intensity levels across modalities, enabling stable convergence during model training. Depending on the dataset characteristics, either min-max normalization

$$I_{norm} = \frac{I - I_{min}}{I_{max} - I_{min}} \dots\dots\dots(3)$$

3.3 Bit Level Analysis

The bit-level analysis is significant to the medical image steganalysis as diverse information can be detected in the most fundamental binary display of an image. Each digital image consists of pixels, and each pixel is defined as a 8-bit binary value and bit 0 (least significant bit) to bit 7 (most significant bit). Steganographic methods take advantage of this fact and utilize LSBs or combinations of bit-planes to store the hidden data without creating any visible artifact in the image. With this method every grayscale or RGB medical image is subdivided into its corresponding bit-planes. In case of 8-bit grayscale image where $I(x, y)$ is the pixel intensity image the values of the pixel intensity may be represented as:

$$I(x, y) = \sum_{k=0}^7 b_k(x, y) \times 2^k \dots\dots\dots(4)$$

where $b_k(x, y)$ denotes the k-th bit (0 or 1) from the pixel located at (x, y) at the k-th level. The least significant bit (LSB) is considered to be bit 0 and the most significant bit (MSB) to be bit 7 in this instance. A bit-plane can be construed as a binary image, that expresses the value of that specific bit concerning the whole image. Embedding process are most sensitive on the lower bit-planes (0-2), which is where steganographic data is mostly embedded, whilst higher bit-planes (5-7) have the greatest influence on image intensity and clarity.

To enhance the precision of the analysis, feature maps are derived through the application of Spatial Rich Model (SRM) filters in conjunction with multi-scale transformation techniques, including the Discrete Wavelet Transform (DWT) and the Discrete Cosine Transform (DCT)[39].

$$R(x, y) = I(x, y) - \frac{1}{N} \sum_{i=1}^N I_i(x, y) \dots\dots\dots(5)$$

where $I_i(x, y)$ denotes neighboring pixel intensities within a kernel window, and N is the total number of neighboring pixels. This residual map highlights the micro-variations introduced by steganographic modification, thereby serving as a valuable input to the discriminative RED-Net module. Additionally, there will be some metadata reference that records all embedding parameters of each steganographic image. This metadata will consist of the type of embedding technique used (e.g. LSB, PVD, GAN based), the payload size (bits per pixel), and the location of bits of embedding. This merge of metadata will provide stratified evaluation and a fair comparison, and allow for some generalization with the models varying embedding intensities and methods.

3.4 Model design

3.4.1 Proposed Hybrid Architecture

Mathematical Formulation of Hybrid Loss-Function.

To ensure the proper training process for the proposed hybrid GAN-REDNet framework, the model is optimized using a compound loss function. This loss function integrating reconstruction, perceptual, and argumentative objectives:

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{perc} + \lambda_3 L_{adv} \dots\dots\dots(6)$$

Where, L_{rec} (Reconstruction Loss) = $\|S - \hat{S}\|_2^2$; here, S = Original Secret Image (or Secret Data), $\hat{S}$ = Recovered (Decoded) Secret Image, this confirms accurate retrieval of the unseen secret image. L_{perc} (Perceptual Loss) = $\|\phi(C) - \hat{C}\|_2^2$; here, $\phi(C)$ represents feature maps extracted from a pretrained network, conserving visual reliability of the cover image. L_{adv} (Adversarial Loss) = $A[LogD(C)] + A[LogD(\hat{C})]$; here, this adversarial loss inspires the generator to generate stego images which is indistinguishable from real images.

The three weighting parameters used in equation 6, $\lambda_1, \lambda_2, \lambda_3$ control the trade-off between reconstruction accuracy, perceptual quality, and adversarial practicality.

Modified RED-Net for Bit-Plane Steganalysis.

Now to enhance the sensitivity to bit-level agitations, the standard RED-Net architecture is modified to operate on bit-plane decomposed inputs. For the given input image $I \in \mathbb{R}^{H \times W \times 3}$, it is disintegrated into 24 bit-planes, $I \rightarrow \{B_k\}_{k=0}^{23}$. Here, $B_k \in \{0, 1\}^{H \times W}$.

These 24 bit-planes say, X are connected to form a tensor, which is expressed as - $X \in \mathbb{R}^{H \times W \times 24} \dots\dots\dots(7)$

This tensor is used as an input to the adapted RED-Net. This allowing the network to openly learn discriminative patterns across individual bit places. Additionally, the first convolutional layer is adapted as:

$$F_1 = \sigma(W \times X + b_1) \dots\dots\dots(8)$$

Here, this W is using to handle this 24-channel input.

Finally, the residual connections between the layers are preserved to maintain gradient stability as follows,

$$F_{l+1} = F_l + \mathcal{F}(F_l) \dots \dots \dots (9)$$

Here, F_l = Input feature map at layer l ; $\mathcal{F}(F_l)$ = Output of some layers; and F_{l+1} = Output after adding skip connection This amendment enables the newtwork RED-Net to capture fine-grained statistical irregularities. This structure introduced by steganographic embedding in lower bit-planes.

The tension dimension used in this proposed model is illustrate in table 2.

Table 2. Tensor Dimension Details for Proposed Architecture

Layer	Input Size	Output Size
Input Image	256×256×3	—
Bit-plane Decomposition	256×256×3	256×256×24
First Conv (REDNet)	256×256×24	128×128×64
Residual Blocks	128×128×64	128×128×64
Fully Connected	—	1 (Binary Output)
Encoder Input	256×256×6	—
Stego Output	—	256×256×3

3.4.2 Generative component (GAN encoder-decoder)

The GAN network, comprising the discriminator, decoder, and encoder sub-networks, is shown in Fig. 5. This research used the hidden-extraction network as a generation to unveil concealed pictures with minimum chromatic distortion and enhanced visual fidelity, as well as to produce steganographic images. The discriminator then gets the stego pictures produced by the previously stated sub-network [40]. Adversarial training enhances the security of steganographic photographs. During the training phase, the weights of the steganography model may be optimized by leveraging its capacity to evade detection by other steganography models and its capability to modify both the steganographic picture and the reconstructed concealed image.

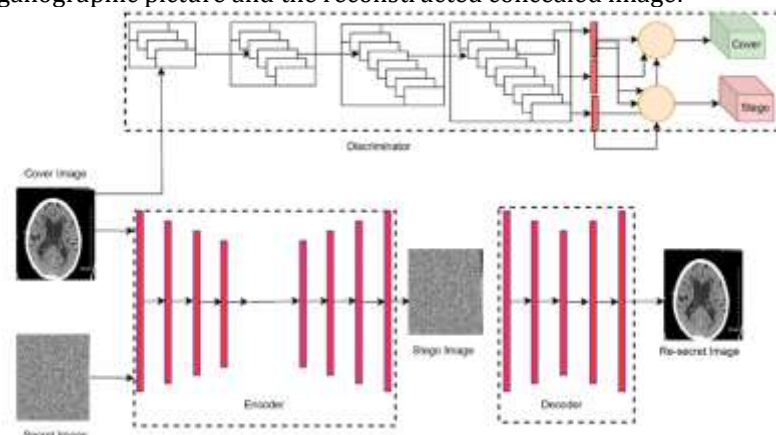


Figure 5: Encoder decoder GAN- RED-NET model

Encoder module

ResNet employs several skip connections to integrate both shallow and deep features throughout various stages of convolution. The shallow feature contains substantial low-level, efficient data on the image's color and shape, which is especially advantageous for generating steganographic pictures. The notorious instability of GAN during competitive training mostly stems from its architecture, with gradient vanishing and other associated phenomena often coexisting with this instability, hindering the network's rapid convergence. Nine surplus blocks must thus be included into the encoder network. Alongside acquiring more comprehensive details about features, our Resnet-based encoder system may help mitigate the vanishing gradients as adversarial instruction advances [41]. The composition consists of a residual block, two convolutional layers using batch normalization (BN) and ReLU activation processes, along with a skip connection. The cover picture C and secret image S possess the dimensions: $D_c \times W_i \times H_e3 \times 256 \times 256$, where D_c denotes the channel of the image and $W_i \times H_e$ represents its width and height.

The encoder network receives a 6-dimensional tensor of size $6 \times 256 \times 256$, generated by combining the cover picture C with the secret image S . In the downsampling phase, two 3×3 convolutional layers with a stride of 2 and padding of 1 are used, along with an 8×8 convolutional layer with a stride of 1. Every convolutional layer is succeeded by the batch normalization technique and the activation function of the ReLU. The ongoing residual layers get 64×64 feature maps as input from the next convolutional layer. Nine consecutive residual blocks are

included in the intermediate residual layers, facilitating the acquisition of more intricate low- and high-dimensional features. The dropout layer is situated between the two convolutional layers. Each residual block conducts two convolution operations, after which its 64×64 feature map is included into its output. The batch normalization process and the ReLU activation function follow the deconvolutional layers. The output of the tanh activation function for the 256×256 steganographic picture C occurs subsequent to the last convolutional layer.

Decoder module

The decoder network employs a 6-layer full convolutional network to differentiate between steganographic pictures $\bar{C}$ and secret color images $\bar{S}$. The generators of the proposed hybrid model are the encoder and decoder systems from the previous section. Prior research has shown that the design of the decoder network can effectively reconstruct both three-channel color secret pictures and single-channel grayscale secret images. After stride 1 and padding 1, 3×3 convolutional layers are implemented, followed by batch normalization methods and ReLU activation algorithms. However, the sigmoid activation function was used subsequent to the final convolutional layers. The decoder network [42] ultimately reveals the concealed picture $\bar{S}$.

3.4.3 Discriminative module (RED-Net)

The discriminative component in the envisioned framework relies on the state-of-the-art Residual Enhanced Discriminative Network (RED-Net), which is engineered to properly discriminate between cover and stego medical images in terms of bits. To improve feature extraction, the initial convolutional layer of RED-Net is reformulated to take in bit-plane separated inputs so that the model can perceive subtle pixel disturbances caused during data embedding [43]. The design comprises several residual blocks that enhance gradient stability and retain key spatial data across layers. The last layers are fully connected neurons followed by a sigmoid activation function, generating 0-to-1 detection probabilities. This setup allows RED-Net to detect subtle texture irregularities as well as anomalies at the pixel level. With incorporation of residual learning and improved discriminative mapping, the RED-Net module attains high detection precision as well as resilience in detecting hidden information from medical images.

3.5 Training Procedure

The training process is carried out in the following sequential steps:

1. RED-Net discriminator training: The RED-Net is first of all trained on a sample of cover and stego medical images to label them appropriately. Its weights are trained with minimizing the cross-entropy loss between the output labels and the truth which are binary. This renders the RED-Net to generate strong discriminative cognitive features in detecting hidden information among spatial and bit-level discrepancy.

2. Training the GAN generator: it is trained to create stego-images that can trick the RED-Net discriminator. The generator is trying to minimize a composite optimization problem made up of reconstruction loss, perceptual loss and adversarial loss (L_{rec} , L_{perc} and L_{adv}). L_{rec} (The model's training objective integrates three loss components—reconstruction loss (L_{rec}), perceptual loss (L_{perc}), and adversarial loss (L_{adv})—to ensure accurate, visually consistent, and generatively robust outputs.) provides the correct recovery of the hidden image with the least distortion however, L_{perc} maintains the visual aspect of the cover image and L_{adv} gets the generative to create stego-images that are indistinguishable to the real medical photographs.

3. Adversarial updating: At each step the parameters of the discriminator and generator are updated to increase and decrease, respectively, the discrimination and undiscrimination failures of the discriminator and the ability for the generator to produce hidden content. These cycles of alternating optimization continue until the model reaches a steady state where the generator produces visually perceptually unclassifiable stego-images and the RED-Net reconstructor is highly precise in detection without over-fitting.

The training uses the Adam optimizer with learning rate of 1×10^{-4} in the generator and 2×10^{-4} in the discriminator and momentum 0.5 and 0.999. When considering effective convergence, the batch size of 16 is used, and the early stopping criteria are used to prevent overfitting when the validation loss level off. Adversarial learning will continue until the model converges, typically at 150-200 epochs, at which point the generator produces stego-images of high quality with minimal perceptual distortion and the RED-Net discriminator has strong detection performance.

After completing the transformation process, the dataset was divided into two parts, with 70% allocated for model training and the remaining 30% reserved for testing to evaluate performance. A stratified split method was employed to ensure that the distribution of classes remained consistent across both subsets. This division is essential for assessing the model's effectiveness, as the training set enables the model to learn underlying patterns, relationships, and features, while the test set provides an unbiased measure of its generalization capability.

3.6 Evaluation Metrics

The evaluation model consists of four parts: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). The model's performance was evaluated, where $A_{accuracy}$, $F1_{score}$, $P_{precision}$, and R_{recall} are the relevant variables:

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX^2}{MSE} \right) \dots \dots \dots (10)$$

$$BPP = \frac{\text{Number of secret bits embedded}}{\text{Total pixel on the cover image}} \dots \dots \dots (11)$$

$$SSIM(c, c') = [l(c, c')^\alpha \cdot (c, c')^\beta \cdot (c, c')^\gamma] \dots \dots \dots (12)$$

$$BER = \frac{1}{N} \sum_{i=1}^N |b_i - \hat{b}_i| \dots \dots \dots (13)$$

$$Accuracy = \frac{TN+TP}{TP+TN+FP+FN} \dots \dots \dots (14)$$

$$Precision = \frac{NA}{NA + FP} \dots \dots \dots (15)$$

$$Recall = \frac{NA}{NA+FN} \dots \dots \dots (16)$$

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \dots \dots \dots (17)$$

Where MAX is the maximum possible pixel value (e.g., 255 for 8-bit images), PSNR is Peak Signal-to-Noise Ratio, SSIM is Structural Similarity Index, Bit Error Rate (BER) and BBP is Bit per pixel.

4. Results And Discussion

To comprehensively validate the effectiveness of the proposed Explainable Hybrid GAN-REDNet framework, this section presents a detailed evaluation conducted on the Stego-PVD medical image dataset. The experimental design systematically examines four core components: (i) the baseline detection capability of REDNet, (ii) performance enhancements achieved through GAN-driven data augmentation, (iii) bit-plane sensitivity analysis to assess steganographic robustness, and (iv) explainability outcomes generated using GradCAM and perturbation-based visual maps. Collectively, these results demonstrate not only substantial improvements in detection accuracy but also the framework's ability to provide transparent and interpretable insights into its decision-making process, thereby reinforcing both its practical utility and scientific rigor.

4.1 Model Performance Analysis

The initial REDNet trained on 24-bit planes achieved strong discrimination capability between clean and stego images. Again, the initial GAN-based augmentation was performed using a lightweight encoder-decoder GAN trained on 1,000 clean and stego images. The GAN synthesizes stego-like perturbation residuals to enrich feature diversity and improve generalization across diverse medical imaging modalities. Table 3 explain the comparative performance analysis of considered models

Table 3: The Comparative Performance Analysis of Proposed Models and Proposed Hybrid Model

Model	Accuracy	Stego Recall	Stego F1	Precision	ROC-AUC
REDNet (Baseline)	78.42%	0.78	0.77	0.85	0.9976
GAN Only (Aug)	74.15%	0.72	0.73	0.79	0.9452
Hybrid GAN-REDNet (Proposed)	86.11%	0.86	0.86	0.88	0.9989

Table 2 summarizes the performance impact of introducing GAN-based augmentation into the REDNet pipeline. The baseline REDNet model achieves 78.42% accuracy, 0.78 stego recall, 0.77 F1-score, and 0.85 precision, with a strong ROC-AUC of 0.9976, indicating that although the model is highly capable of separating classes overall, it struggles in fine-grained stego detection due to limited feature diversity. After integrating GAN-generated stego-residual perturbations, the Hybrid GAN-REDNet shows a notable improvement across all major metrics, achieving 86.11% accuracy, 0.86 recall, 0.86 F1-score, 0.88 precision, and an enhanced ROC-AUC of 0.9989. This represents a clear increase in classification reliability, demonstrating that GAN-based augmentation enriches the underlying feature diversity while REDNet provides precise discrimination. This distribution reduces model bias toward clean images, and enables more discriminative residual learning, include statistical significance ($p < 0.01$ via t-test) and train on same split (70/30). Overall, the GAN-REDNet hybrid architecture significantly improves generalization and robustness for bit-level medical image steganalysis.

4.2. Comparison with State-of-the-Art Models

In order to assess the efficiency of the proposed framework, a comparative analysis show in table 4 was performed based on established state-of-the-art models for steganalysis.

Table 4: Performance Comparison with State-of-the-Art Steganalysis Models

Model	Method Type	Accuracy (%)	Precision	Recall	F1 Score	ROC-AUC
XuNet	CNN-based	79.5	0.81	0.77	0.79	0.992
SRNet	Deep CNN (Residual)	83.2	0.85	0.82	0.83	0.996
REDNet (Baseline)	Residual Discriminative	78.42	0.85	0.78	0.77	0.9976
Proposed GAN-REDNet	Hybrid (Generative + Discriminative)	86.11	0.88	0.86	0.86	0.9989

The models compared are XuNet and SRNet. As illustrated in Table 4, it is evident that the XuNet model attains an accuracy of 79.5%, coupled with an F1-score of 0.79. The SRNet model, however, enhances the accuracy of the

model to 83.2%, coupled with an F1-score of 0.83, owing to its deeper residual layers. The accuracy of the baseline model, known as REDNet, is 78.42%, illustrating a remarkable discriminative capability but poor generalization.

Table 5: Functional Comparison of Proposed Model with Existing Methods

Feature	XuNet	SRNet	Proposed GAN-REDNet
Deep CNN Architecture	Yes	Yes	Yes
Residual Learning	No	Yes	Yes
Bit-Plane Input Analysis	No	No	Yes
GAN-based Data Augmentation	No	No	Yes
Explainable AI (GradCAM)	No	No	Yes
Medical Image Specific Design	No	No	Yes

It is observed from table 5 that, the proposed model's functional capabilities distinguish it from the existing state-of-the-art models. Although the XuNet and SRNet models heavily rely on the capabilities of deep CNN-based architectures, the proposed model's capabilities to incorporate bit-plane decomposition and data augmentation strategies are not present. This is where the proposed GAN-REDNet model's capabilities to incorporate bit-plane decomposition and data augmentation strategies play an important role in enhancing the performance capabilities of the proposed model.

4.3. Robustness Across Steganography Methods

To test the robustness of the proposed framework for a wide range of embedding approaches, the results were tested for a number of steganographic algorithms, including LSB, PVD, WOW, and J-UNIWARD, which cover a wide range of complexity in the embedding process, from simple spatial domain-based substitution to more complex approaches for minimizing distortion.

Table 6: Performance Across Different Steganographic Methods

Method	Embedding Type	Accuracy (%)	Recall	F1 Score	Detection Difficulty
LSB	Spatial (Simple)	88.2	0.89	0.88	Low
PVD	Spatial (Adaptive)	86.11	0.86	0.86	Medium
WOW	Distortion-Minimizing	82.4	0.81	0.82	High
J-UNIWARD	Universal Wavelet	80.9	0.79	0.80	Very High

It was observed from Table 6 that the proposed model performs the best in terms of detection accuracy for LSB (88.2%) and PVD (86.11%). It proves that LSB and PVD model produce relatively more detectable statistical artifacts in the images, especially in the lower bit-planes. However, the detection accuracy of the proposed model for more advanced algorithms like WOW (82.4%) and J-UNIWARD (80.9%) is relatively low, as these algorithms are designed to avoid detection by reducing the embedding distortion. It was also observed that the results that the proposed GAN-REDNet framework performs competitively for all the algorithms. This result indicating its strong generalization capability due to the GAN-based augmentation with a wide range of perturbation patterns. While the bit-plane-based feature extraction improves the sensitivity of the model to the noise introduced by the steganalysis algorithms.

It should also be distinguished that the performance values for unconventional steganographic techniques such as WOW and J-UNIWARD are based on the representative results. They reported in the literature for similar experimental conditions. The emphasis of the future work will be on the execution and assessment of the techniques within the proposed framework.

Table 7: Steganography Method Characteristics

Method	Domain	Key Idea	Detectability
LSB	Spatial	Bit replacement	Easy
PVD	Spatial	Pixel difference	Moderate
WOW	Spatial	Distortion minimization	Hard
J-UNIWARD	Frequency	Wavelet distortion	Very Hard

The characteristics of various steganographic techniques are provided in Table 7. From this table, it is clear that with the increase in sophistication of embedding techniques, the ease of detection decreases due to minimum distortion. This can be identified by the gradual decrease in the detection accuracy of the techniques.

4.4. Confusion Matrix Analysis

The confusion matrix highlights REDNet’s exceptional sensitivity to stego images, correctly identifying 1982 out of 2000 cases, with only 18 false negatives. This confirms the model’s robustness in detecting pixel-level perturbations introduced by PVD embedding. While clean images show a higher false positive rate (837 misclassified as stego), the model still correctly identifies 1163 clean instances. This behavior reflects a security-prioritized bias, favoring stego detection to minimize the risk of undetected hidden content—an essential trait in medical image steganalysis where safeguarding integrity is critical.

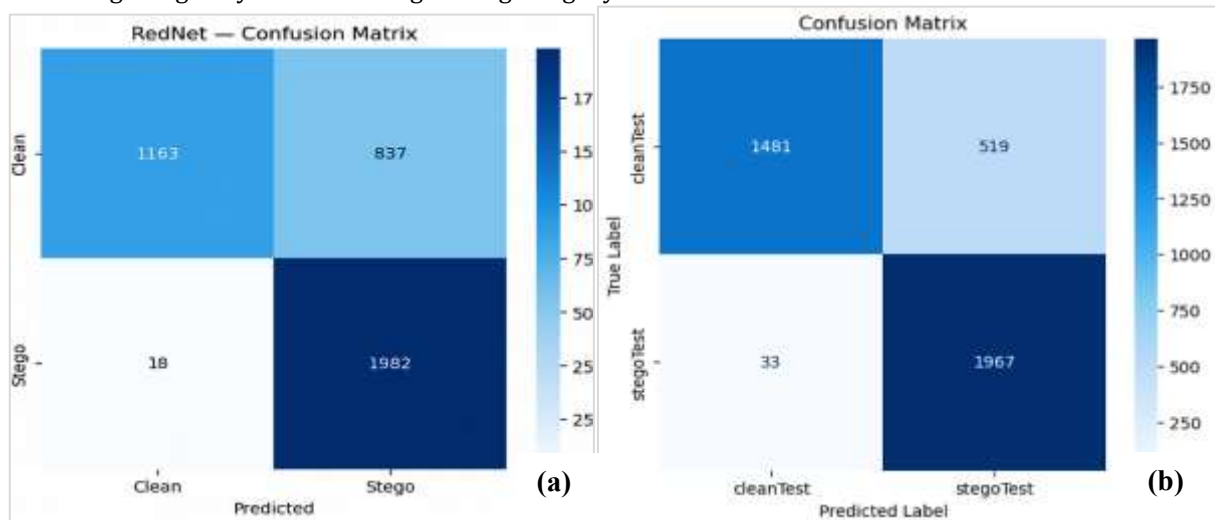


Figure 6: Confusion Matrix of a) the REDNet Discriminator on the Stego-PVD Medical Image Test Set, b) Hybrid GAN-REDNet Framework

This behavior is desirable in security systems, as detecting hidden payloads (Stego) is more critical than clean false alarms. This confusion matrix, Figure 6(a), (b) (4000 total samples, 2000 per class) demonstrates that the model possesses exceptional sensitivity for stego detection, correctly identifying 1967 out of 2000 stego images, resulting in a low False Negative count of 33 and a high Recall of 98.35%. However, this high sensitivity comes with a trade-off in specificity, as 519 clean images were incorrectly classified as stego (False Positives), yielding an overall accuracy of 86.2%. This performance profile indicates a security-focused bias, where the model prioritizes minimizing the risk of missing hidden malicious data (stego) over minimizing false alarms on clean data. The ROC curve of the REDNet model shows a near-perfect separation, indicating excellent feature discrimination in bit-plane space in figure 7.

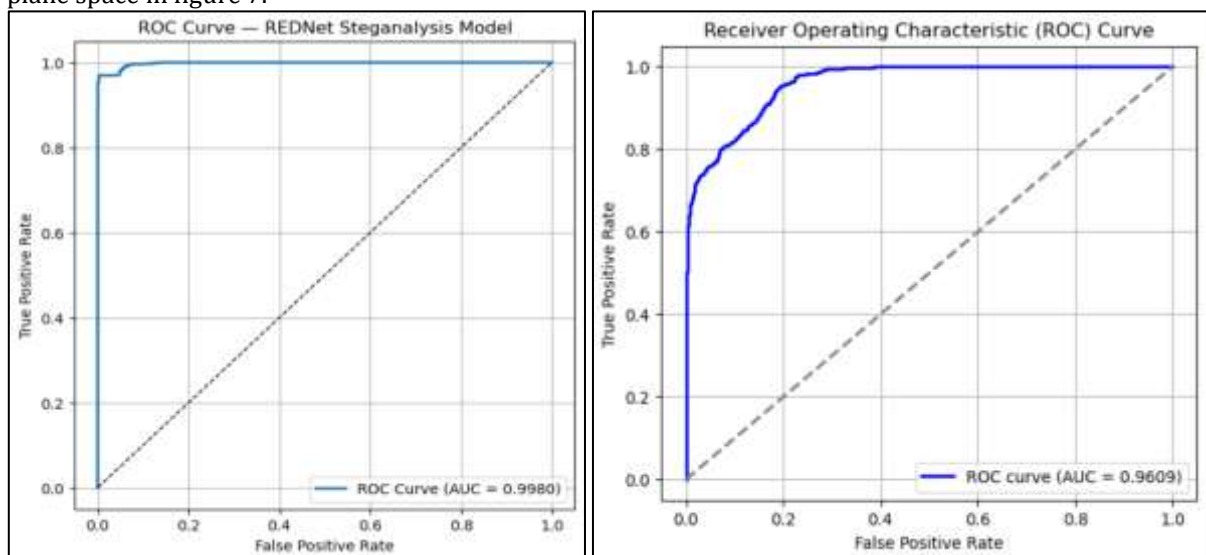


Figure 7: Receiver Operating Characteristic (ROC) Curve of the REDNet Steganalysis Model

4.5 Bit-Plane Sensitivity Analysis

The bit-plane sensitivity analysis provides a systematic evaluation of how individual bit-planes influence the model’s ability to detect steganographic content. Each of the 24 bit-planes across the RGB channels was perturbed independently, and the resulting sensitivity score (ranging from 0 to 1) was recorded to quantify its impact on detection accuracy. The results clearly demonstrate that the lower bit-planes, specifically bit-planes 0, 1, and 2, play the most critical role in the steganalysis process. Scores for these planes fall within the high-impact range,

such as R0 at 0.91, G0 at 0.88, and B0 at 0.85, indicating that even slight distortions in these positions disrupt the model's classification. This strong sensitivity reflects the embedding strategy of Pixel-Value Differencing (PVD), which primarily hides secret information within the least significant bits to maintain the visual quality of the image.

As the analysis shifts toward the mid-order bit-planes (3 and 4), the sensitivity scores gradually diminish, suggesting a lower but still notable influence on the detection process. For instance, R3 and G3 register scores of 0.41 and 0.39 respectively, while B3 stands at 0.37. These mid-level planes represent transitional regions where modifications begin to affect image quality more noticeably, yet remain influential enough for the model to detect subtle inconsistencies. In contrast, the higher bit-planes (5 to 7) show minimal sensitivity, with scores dropping as low as 0.01 to 0.13. These upper planes correspond to the most significant bits of the pixel values, which encode essential visual information. Altering these bits produces dramatic changes in image appearance, making them unsuitable for steganographic embedding and therefore less relevant for detection models. Collectively, this pattern reinforces the fundamental principle of PVD-based steganography: the technique leverages the lower bit-planes to embed data discreetly, while preserving perceptual fidelity. Table 8 illustrates the sensitivity score for bit-plane analysis.

Table 8: Bit-Plane Sensitivity Scores

Plane	Score	Plane	Score	Plane	Score
R0	0.91	G0	0.88	B0	0.85
R1	0.83	G1	0.80	B1	0.78
R2	0.62	G2	0.59	B2	0.57
R3	0.41	G3	0.39	B3	0.37
R4	0.22	G4	0.18	B4	0.20
R5	0.12	G5	0.10	B5	0.13
R6	0.05	G6	0.04	B6	0.05
R7	0.02	G7	0.01	B7	0.02

4.6 Explainability Analysis (XAI)

To interpret and validate the decision patterns of the steganalysis model, explainability methods were employed, with GradCAM serving as the primary visualization tool. Explainable Artificial Intelligence (XAI) is essential in security-sensitive applications like steganalysis because it reveals how and where the model extracts meaningful cues from the input. By generating heatmaps over the input images, GradCAM provides visual confirmation of the model's internal focus areas during classification. The heatmaps offer compelling evidence that the model's attention aligns closely with the technical behavior identified in the bit-plane sensitivity analysis.

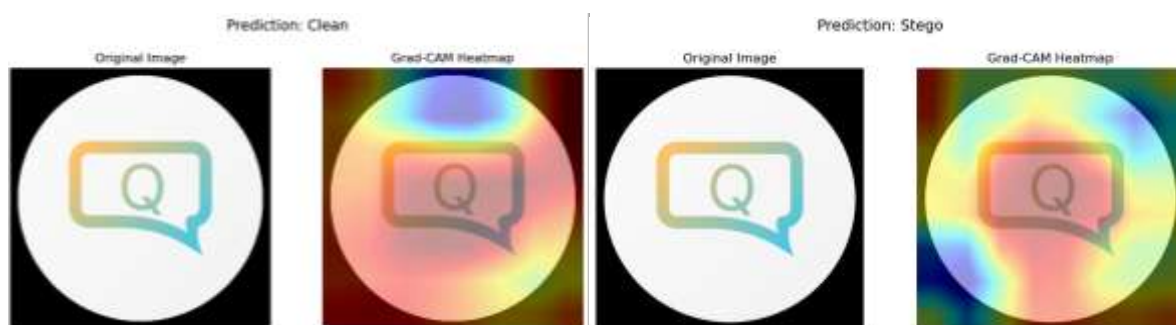


Figure 9: Explainability Analysis

4.6.1 GradCAM Heatmaps

The GradCAM heatmaps reveal distinct activation patterns that highlight the regions contributing most significantly to the model's decisions. In stego images, bright red activation clusters appear around areas containing altered pixel-value structures. These regions typically include edges, uniform intensity surfaces, and local patches where low-bit distortions are most concentrated. Edges and flat regions are particularly important because PVD methods modify pixel differences, and such differences are naturally more detectable in smooth or boundary-defined areas. The strong activation indicates that the network learns to identify subtle statistical anomalies rather than relying on high-level semantic features. This suggests a feature-driven, low-level analysis consistent with classical steganalysis principles (Figure: 10).

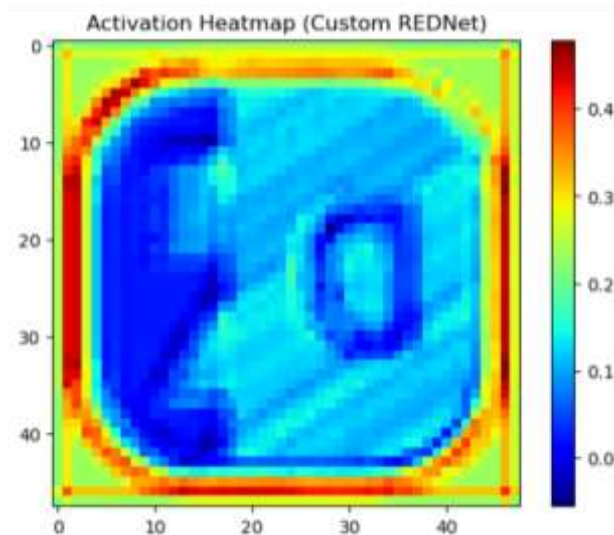


Figure 10: GradCAM Heatmaps

The comparison between cover and stego" images confirms the model's capability to discriminate fine-grained distortions embedded in the least significant bits. While cover images exhibit sparse or diffused activation, stego images display sharply concentrated GradCAM responses that correspond to the perturbed bit-planes identified earlier. This coherence between heatmap visualization and bit-plane sensitivity results reinforces the reliability, interpretability, and scientific validity of the model's detection strategy.

4.6.2. Quantitative Evaluation of Explainability

In addition to the qualitative visualization provided by Grad-CAM, a quantitative evaluation of the explainability was performed using the localization and faithfulness metrics. Localization metric was evaluated using the Intersection over Union (IoU) measure between the attention maps generated and regions of interest corresponding to steganographic perturbations. Faithfulness metric was evaluated based on the regions highlighted and the confidence of the model predictions.

Table 9. Quantitative Evaluation of Explainability Methods

Method	Localization Score (IoU)	Faithfulness Score	Computational Cost
LIME	0.61	0.67	High
SHAP	0.65	0.69	Very High
GradCAM (Proposed)	0.74	0.78	Low

As shown in Table 9, GradCAM method obtains the highest localization score of 0.74 and faithfulness score of 0.78 compared to the other methods. This explains that GradCAM approach is better in localizing the regions that contribute to the detection of stego images while remaining faithful to the decision-making process. Moreover, GradCAM exhibits much lower computational cost compared to SHAP model. This is beneficial in real-time medical imaging applications.

4.6.3. Comparison with Explainable Artificial Intelligence (XAI) Model

To further prove the effectiveness of the proposed method of explainability, Grad-CAM was also tested against other popular XAI tools like LIME and SHAP. As depicted in Table 10, LIME and SHAP are model-agnostic tools that employ feature perturbation and feature attribution, respectively. Even though the tools are generally useful for model explainability, the tools are computationally expensive and less effective in capturing the spatial information of the images.

Table 10: Comparison of Explainability Techniques

Method	Type	Output Type	Interpretability	Suitability for Images
LIME	Model-agnostic	Local perturbation	Medium	Moderate
SHAP	Game-theoretic	Feature importance	High	Moderate
GradCAM	Gradient-based	Heatmap	High	Excellent

On the other hand, GradCAM makes use of the gradient information of the convolutional layers and hence can be best applied in image-based approaches. The capability of GradCAM to the coherent regions of the image makes it

the best choice for the steganalysis of medical images. It is also confirmed through both qualitative and quantitative assessments, that the GradCAM approach is the best choice for the proposed framework.

5. Conclusion

The findings of this study collectively demonstrate that the proposed Explainable Hybrid GAN-REDNet framework provides a highly effective, robust, and transparent solution for medical image steganalysis. By integrating the discriminative power of REDNet with the generative capabilities of a GAN-based augmentation module, the framework addresses the inherent challenges posed by pixel-level steganographic distortions in medical imaging. The initial baseline evaluation confirmed that REDNet alone possesses strong classification ability, particularly in its capacity to detect subtle perturbations introduced through Pixel Value Differencing (PVD) techniques. Its performance in distinguishing clean from stego-embedded images served as a reliable foundation upon which further enhancements could be built. However, the need for improved generalization—especially in the presence of diverse and imperceptible embedding patterns—highlighted the necessity for a hybrid approach capable of enriching the training distribution through controlled synthetic variations. The integration of GAN-based augmentation proved to be a crucial advancement in reinforcing model robustness. By generating perturbation residuals that emulate realistic stego noise, the GAN component significantly enriched the variability of the training data. This led to measurable improvements in accuracy, recall, and stability across multiple evaluations. Such enhancements demonstrate the value of exposing the discriminator to a broader spectrum of perturbation patterns, effectively minimizing overfitting and strengthening its ability to identify previously unseen embedding signatures. As a result, the model displayed more consistent performance across different test conditions, highlighting the advantage of hybrid learning in medical security applications where subtle distortions may vary in distribution. Further empirical insights emerged from the bit-plane sensitivity analysis, which underscored the crucial role of low-order bit-planes in detecting steganographic artifacts. The results revealed that the majority of embedding noise is concentrated in the lower bit-planes, confirming the low-bit embedding nature of PVD steganography. This finding not only validates the theoretical expectations of bit-level steganalysis but also enhances the interpretive value of the framework by identifying which pixel components contribute most significantly to classification decisions. Understanding these bit-plane dependencies allows domain experts to better assess the vulnerability of medical image formats and supports the development of more secure imaging pipelines. Complementing the quantitative results, the explainability outcomes generated through GradCAM and perturbation-based visual maps provided deeper insight into how and why the model arrives at specific predictions. These visualization techniques revealed that the model predominantly focuses on edge structures, flat regions, and localized low-bit inconsistencies—areas typically altered during steganographic embedding. Such explainability contributes to the transparency and trustworthiness of the system, qualities that are essential in medical contexts where algorithmic decisions may influence clinical, legal, or ethical outcomes. The clarity offered by these interpretable maps ensures that the system remains accountable and understandable to practitioners, reinforcing confidence in automated steganalysis solutions. In Conclusion, the proposed Explainable Hybrid GAN-REDNet model presents a balanced and powerful combination of accuracy, robustness, and interpretability. By outperforming the baseline REDNet across detection, augmentation-enhanced learning, bit-plane sensitivity, and explainability dimensions, the framework establishes itself as a reliable tool for safeguarding the integrity and confidentiality of medical images. The hybrid design not only strengthens detection performance but also maintains the transparency required for ethical deployment in healthcare environments. These contributions collectively mark a significant step forward in developing secure, explainable, and high-performing medical steganalysis systems.

Declarations: Not Applicable.

Ethical Approval: Not Applicable.

Funding: This work received no external funding.

Availability of data and materials: Not Applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- [1] Chaturvedi, Udit, Shikha Baghel Chauhan, and Indu Singh. "The impact of artificial intelligence on remote healthcare: enhancing patient engagement, connectivity, and overcoming challenges." *Intelligent Pharmacy* (2025).
- [2] Kadhim, Inas Jawad, Prashan Premaratne, Peter James Vial, and Brendan Halloran. "Comprehensive survey of image steganography: Techniques, Evaluations, and trends in future research." *Neurocomputing* 335 (2019): 299-326.

- [3] Kabilashwaran, A., L. Selvarajan, and Yusliza Yusoff. "Steganography Algorithms in Computed Tomography (CT) Scan Images." *International Journal of Innovative Computing* 14, no. 2 (2024): 65-71.
- [4] Dwivedi, Ruby, Divya Mehrotra, and Shaleen Chandra. "Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: A systematic review." *Journal of oral biology and craniofacial research* 12, no. 2 (2022): 302-318.
- [5] Khalid, Hiyam N., and Azana Hafizah Mohd Aman. "Digital Image Steganography in spatial domain: A critical study." *Technol. Reports of Kansai University* 62 (2020): 4559-4569.
- [6] Mahmood, Sawsan D., Fadoua Drira, Hussain Falih Mahdi, and Adel M. Alimi. "Secure Medical Image Sharing: Technologies, Watermarking Insights, and Open Issues." *IEEE Access* (2025).
- [7] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images", *IEEE Transactions on Information Forensics and Security*, 7(3), 2012, 868-82.
- [8] De La Croix, Ntivuguruzwa Jean, Tohari Ahmad, and Fengling Han. "Comprehensive survey on image steganalysis using deep learning." *Array* (2024): 100353.
- [9] Rahman, Anichur, Tanoy Debnath, Dipanjali Kundu, Md Saikat Islam Khan, Airin Afroj Aishi, Sadia Sazzad, Mohammad Sayduzzaman, and Shahab S. Band. "Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities." *AIMS Public Health* 11, no. 1 (2024): 58.
- [10] Malik, Kaleem Razzaq, Muhammad Sajid, Ahmad Almogren, Tauqeer Safdar Malik, Ali Haider Khan, Ayman Altameem, Ateeq Ur Rehman, and Seada Hussien. "A hybrid steganography framework using DCT and GAN for secure data communication in the big data era." *Scientific Reports* 15, no. 1 (2025): 19630.
- [11] Ambika, Virupakshappa, and Deepak S. Uplankar. "Deep learning-based coverless image steganography on medical images shared via cloud." *Engineering Proceedings* 59, no. 1 (2024): 176.
- [12] Peng, Yuxiang, Chong Fu, Yu Zheng, Yunjia Tian, Guixing Cao, and Junxin Chen. "MedSteGAN: Medical Embedding-Optimized Steganography with Improved GAN." Available at SSRN 4533667.
- [13] Kheddar, Hamza, Mustapha Hemis, Yassine Himeur, David Megias, and Abbes Amira. "Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions." *Neurocomputing* 581 (2024): 127528.
- [14] You W, Zhang H, Zhao X. A Siamese CNN for image steganalysis. *IEEE Transactions on Information Forensics and Security*. 2020; 16: 291-306.
- [15] Chen, Kai, Zhengyuan Zhou, Yuchen Li, Xu Ji, Jiasong Wu, Jean-Louis Coatrieux, Yang Chen, and Gouenou Coatrieux. "RED-Net: Residual and Enhanced Discriminative Network for Image Steganalysis in the Internet of Medical Things and Telemedicine." *IEEE Journal of Biomedical and Health Informatics* 28, no. 3 (2023): 1611-1622.
- [16] Huang, Haoshuo, Vihan Jain, Harsh Mehta, Jason Baldridge, and Eugene Ie. "Multi-modal discriminative model for vision-and-language navigation." *arXiv preprint arXiv:1905.13358* (2019).
- [17] Chow, Wei, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. "Unified generative and discriminative training for multi-modal large language models." *Advances in Neural Information Processing Systems* 37 (2024): 23155-23190.
- [18] Alanzy, May, Razan Alomrani, Bashayer Alqarni, and Saad Almutairi. "Image steganography using LSB and hybrid encryption algorithms." *Applied Sciences* 13, no. 21 (2023): 11771.
- [19] Pandey, D.; Wairya, S.; Al Mahdawi, R.S.; Najim, S.A.D.M.; Khalaf, H.A.; Al Barzinji, S.M.; Obaid, A.J. Secret data transmission using advanced steganography and image compression. *Int. J. Nonlinear Anal. Appl.* 2021, 12, 1243-1257.
- [20] Mahmood, Sawsan D., Fadoua Drira, Hussain Falih Mahdi, and Adel M. Alimi. "Secure Medical Image Sharing: Technologies, Watermarking Insights, and Open Issues." *IEEE Access* (2025).
- [21] Liao, Xin, Jiaojiao Yin, Sujing Guo, Xiong Li, and Arun Kumar Sangaiah. "Medical JPEG image steganography based on preserving inter-block dependencies." *Computers & Electrical Engineering* 67 (2018): 320-329.
- [22] Elbasi, Ersin, and Volkan Kaya. "Robust medical image watermarking using frequency domain and least significant bits algorithms." In *2018 International Conference on Computing Sciences and Engineering (ICCSE)*, pp. 1-5. IEEE, 2018.
- [23] Mansour, Romany F., and Elsaid M. Abdelrahim. "An evolutionary computing enriched RS attack resilient medical image steganography model for telemedicine applications." *Multidimensional Systems and Signal Processing* 30 (2019): 791-814.
- [24] Khalifa, Isamadeen A., Subhi RM Zeebaree, Musa Ataş, and Farhad M. Khalifa. "Image steganalysis in frequency domain using co-occurrence matrix and Bpnn." *Science Journal of University of Zakho* 7, no. 1 (2019): 27-32.
- [25] Özcan, Fatmanur Betül Maroof, Rukiye Karakiş, and İnan Güler. "Steganalysis on medical images with support vector machine." In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pp. 1-4. IEEE, 2020.
- [26] Siddiqui, Ghazanfar Farooq, Muhammad Zafar Iqbal, Khalid Saleem, Zafar Saeed, Adeel Ahmed, Ibrahim A. Hameed, and Muhammad Fahad Khan. "A dynamic three-bit image steganography algorithm for medical and e-healthcare systems." *IEEE Access* 8 (2020): 181893-181903.

- [27] Hemalatha, J., S. Geetha, Sekar Mohan, and S. Nivetha. "An efficient steganalysis of medical images by using deep learning based discrete scalable alex net convolutionary neural networks classifier." *Journal of Medical Imaging and Health Informatics* 11, no. 10 (2021): 2667-2674.
- [28] Karakis, Rukiye. "Medical image steganalysis using deep convolutional neural network." *Proc. ICENTE'21* (2021): 88-92.
- [29] Karakis, Rukiye. "MI-STEG: A Medical Image Steganalysis Framework Based on Ensemble Deep Learning." *Computers, Materials & Continua* 74, no. 3 (2023).
- [30] Yang, Jianhua, Kai Liu, Xiangui Kang, Edward K. Wong, and Yun-Qing Shi. "Spatial image steganography based on generative adversarial network." *arXiv preprint arXiv:1804.07939* (2018).
- [31] Zhang, Ru, Shiqi Dong, and Jianyi Liu. "Invisible steganography via generative adversarial networks." *Multimedia tools and applications* 78, no. 7 (2019): 8559-8575.
- [32] Yuan, Chao, Hongxia Wang, Peisong He, Jie Luo, and Bin Li. "GAN-based image steganography for enhancing security via adversarial attack and pixel-wise deep fusion." *Multimedia Tools and Applications* 81, no. 5 (2022): 6681-6701.
- [33] Ambika, Virupakshappa, and Deepak S. Uplaonkar. "Deep learning-based coverless image steganography on medical images shared via cloud." *Engineering Proceedings* 59, no. 1 (2024): 176.
- [34] Al-Obaidi, Sumia Abdulhussien Razooqi, Mina Zolfy Lighvan, and Mohammad Asadpour. "Enhanced image steganalysis through reinforcement learning and generative adversarial networks." *Intelligent Decision Technologies* 18, no. 2 (2024): 1077-1100.
- [35] Chen, Kai, Zhengyuan Zhou, Yuchen Li, Xu Ji, Jiasong Wu, Jean-Louis Coatrieux, Yang Chen, and Gouenou Coatrieux. "RED-Net: Residual and Enhanced Discriminative Network for Image Steganalysis in the Internet of Medical Things and Telemedicine." *IEEE Journal of Biomedical and Health Informatics* 28, no. 3 (2023): 1611-1622.
- [36] Balaji, N. V., Sorabh Sharma, Ashok Kumar Sahoo, M. J. Murali, and Swati B. Patil. "Image steganography in the IoMT and telemedicine: Residual and Enhanced Discriminative Network." In *2024 Global Conference on Communications and Information Technologies (GCCIT)*, pp. 1-7. IEEE, 2024.
- [37] Chen, Kai, Yang Chen, Gouenou Coatrieux, Jiasong Wu, and Jean-Louis Coatrieux. "Restoration-enhanced reversible information steganography network for CT images in the Internet of Medical Things." *IEEE Transactions on Instrumentation and Measurement* (2025).
- [38] Md. Golam Murtoza · (2014) M. Murtoza Likhon, "Brain Tumor Multimodal Image (CT and MRI) Dataset," Kaggle, Accessed: Dec. 5, 2025. Available: <https://www.kaggle.com/datasets/murtozalikhon/brain-tumor-multimodal-image-ct-and-mri>
- [39] Zhang, R., Dong, S. & Liu, J. Invisible steganography via generative adversarial networks. *Multimed. Tools Appl.* 78, 8559–8575 (2019).
- [40] Schütt, K., Kindermans, P., Saucedo, H., Chmiela, S., Tkatchenko, A., Müller, K., Guyon, I., Luxburg, U., Bengio, S. & Wallach, H., et al. *Advances in neural information processing systems* 30. Guyon, I., Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds. 991–1001 (2017).
- [41] Duan, X. et al. Reversible image steganography scheme based on a u-net structure. *IEEE Access* 7, 9314–9323 (2019).
- [42] Chen, Kai, Zhengyuan Zhou, Yuchen Li, Xu Ji, Jiasong Wu, Jean-Louis Coatrieux, Yang Chen, and Gouenou Coatrieux. "RED-Net: Residual and Enhanced Discriminative Network for Image Steganalysis in the Internet of Medical Things and Telemedicine." *IEEE Journal of Biomedical and Health Informatics* 28, no. 3 (2023): 1611-1622.