XViT-GAT: Explainable Vision Transformer–Graph Attention Network Framework for Robust Malware Detection in Steganographic Images

Celia Anthony
Xavier Institute of Computer Applications, Ahmedabad, Gujarat, PhD Scholar, Faculty of Computer Applications, Ganpat University, Ganpat Vidyanagar, Gujarat, India

Savan Patel
Faculty of Computer Applications, Acharya Motibhai Patel Institute of Computer Studies (AMPICS), Ganpat University, Ganpat Vidyanagar, Gujarat, India

Nirbhay Kumar Chaubey
Faculty of Computer Applications, Acharya Motibhai Patel Institute of Computer Studies (AMPICS), Ganpat University, Ganpat Vidyanagar, Gujarat, India

ABSTRACT
Image steganalysis is a critical component of digital forensics and cybersecurity, as it is capable of identifying concealed information that is embedded within digital images. The capacity of image steganalysis systems to identify sophisticated steganographic methods has been substantially enhanced by recent advancements in artificial intelligence. Moreover, conventional feature learning methods and existing augmentation-driven approaches are inadequate for security-critical applications due to their limited interpretability and absence of robust generalization capability. This research suggests the development of XViT-GAT: Explainable Vision Transformer–Graph Attention Network Framework for Robust Malware Detection in Steganographic Images, an advanced hybrid deep learning framework that integrates self-supervised representation learning, vision transformer-based feature extraction, graph attention-based artefact modelling, and explainable artificial intelligence, in order to address these limitations. Initially, the adaptive noise extraction module is implemented in the proposed framework to enhance the subtle steganographic residual patterns and suppress irrelevant image information. The hybrid Vision Transformer encoder effectively captures both local steganographic irregularities and long-range dependencies by processing the extracted representations through multi-head self-attention mechanisms. Additionally, complex relationships among spatial artefacts, frequency-domain variations, compression inconsistencies, and malware-related embedding characteristics are modelled using a Graph Attention Network (GAT) module. The detection capability against sophisticated and adaptive malware-hiding techniques is enhanced by this graph-based feature reasoning. The XViT-GAT framework that has been proposed is intended to enable the precise, dependable, and comprehensible identification of malware that is concealed within steganographic images. The anticipated results indicate that the proposed method can surpass the constraints of conventional CNN-based steganalysis, GAN-based augmentation methods, and reinforcement learning-assisted detection systems by achieving enhanced zero-day malware detection capability, reliable decision explanations, and improved feature generalization.

Keywords
Malware Detection, Steganographic Images; Vision Transformer, Graph Attention Network, Explainable Artificial Intelligence, Self-Supervised Learning, Steganalysis, Deep Learning, Cybersecurity, Image-Based Malware Analysis, Contrastive Learning, Zero-Day Malware Detection.

1. INTRODUCTION
The rapid evolution of cyber threats has led attackers to adopt increasingly sophisticated concealment strategies, including steganography, in which malicious payloads, executable fragments, commands, or configuration information are covertly embedded within apparently benign digital images. This emerging class of stegomalware is particularly challenging because conventional signature-based and heuristic security mechanisms may fail to identify hidden malicious content without altering or deeply analyzing the carrier image [1]. Recent research has therefore shifted toward deep-learning-based steganalysis, which can automatically learn subtle statistical and spatial artifacts introduced by information-hiding operations [2]. At the same time, visualization-based malware analysis has gained attention by representing executable or binary information as images and applying advanced deep neural architectures for classification. In particular, Vision Transformers (ViTs) have demonstrated considerable potential because their self-attention mechanisms can capture long-range dependencies and global structural relationships that may be difficult for conventional convolutional models to represent [3], [4]. Recent ViT-based malware studies further demonstrate the usefulness of regional/local attention and image-level representations for distinguishing sophisticated malware patterns [5].

Nevertheless, image-oriented analysis alone may inadequately represent the relational dependencies among suspicious regions, embedded components, structural artifacts, and malware-related features. Graph-based learning provides a complementary mechanism for modeling these relationships. Graph Attention Networks (GATs) dynamically assign importance to neighboring nodes, allowing security models to emphasize highly informative structural associations rather than treating all relationships equally. Recent malware research has demonstrated the effectiveness of GAT-based representations for combining structural and semantic information, while newer graph-informed Transformer architectures indicate that integrating Transformer representations with graph attention can improve contextual malware modeling [6], [7]. Moreover, explainability remains an important requirement because highly accurate deep models can still operate as black boxes, limiting analysts' ability to understand why a sample has been classified as malicious. Recent studies have consequently incorporated explainable artificial intelligence (XAI) into malware analysis to identify influential features and malicious regions, and graph-attention-based approaches have demonstrated the ability to localize malicious code through interpretable attention mechanisms [8], [9].

To address these challenges, this study proposes XViT-GAT: an Explainable Vision Transformer–Graph Attention Network framework for robust malware detection in steganographic images. The proposed framework combines the complementary capabilities of ViT and GAT, where the ViT learns discriminative global and patch-level visual representations from steganographic images, while the GAT models relationships among the extracted suspicious regions/features to capture their contextual and structural dependencies. An explainability layer is further incorporated to expose the image regions, patches, or graph components that contribute strongly to malware predictions, thereby providing security analysts with interpretable evidence rather than only a classification label. Through this integrated visual–relational–explainable learning strategy, XViT-GAT is designed to improve the detection of concealed malware while strengthening robustness, interpretability, and practical applicability against evolving stegomalware threats.

1.1 Main Contribution of this research

- To develop and suggests the development of XViT-GAT: Explainable Vision Transformer–Graph Attention Network Framework for Robust Malware Detection in Steganographic Images, an advanced hybrid deep learning framework that integrates self-supervised representation learning, vision transformer-based feature extraction, graph attention-based artefact modelling, and explainable artificial intelligence, in order to address these limitations.
- Initially, the adaptive noise extraction module is implemented in the proposed framework to enhance the subtle steganographic residual patterns and suppress irrelevant image information.
- Finally, the hybrid Vision Transformer encoder effectively captures both local steganographic irregularities and long-range dependencies by processing the extracted representations through multi-head self-attention mechanisms.

The rest of this article is organized as follows: section 2 provides a literature survey, proposed method are explained in section 3, result and discussion are explained in section 4, and finally conclusion and future work are explained in section 5.

2. LITERATURE SURVEY

Kuznetsov et al. [10] proposed an extensive investigation into the application of artificial intelligence, specifically Convolutional Neural Networks (CNNs), in image steganography detection. We initially evaluated the state-of-the-art steganalysis model, SRNet, on various image steganography techniques, including WOW, HILL, S-UNIWARD, and the innovative Spread Spectrum Image Steganography (SSIS). We found SRNet's performance on SSIS detection to be lower compared to other methods, prompting us to fine-tune the model using SSIS datasets. Subsequent experiments showed significant improvement in SSIS detection, albeit at the cost of minor performance degradation as to other techniques. Our findings underscore the potential and adaptability of AI-based steganalysis models. However, they also highlight the need for a delicate balance in model adaptation to maintain effectiveness across various steganography techniques.

Płachta et al. [11] proposed an the performance of employing various shallow and deep learning algorithms in image steganography detection is analyzed. The data, images from the BOSS database, were used with information hidden using three popular steganographic algorithms: JPEG universal wavelet relative distortion (J-Uniward), nsF5, and uniform embedding revisited distortion (UERD) at two density levels. Various feature spaces were verified, with the discrete cosine transform residuals (DCTR) and the Gabor filter residuals (GFR) yielding best results. Almost perfect detection was achieved for the nsF5 algorithm at 0.4 bpnzac density (99.9% accuracy), while the detection of J-Uniward at 0.1 bpnzac density turned out to be hardly possible (max. 56.3% accuracy). The ensemble classifiers turned out to be an encouraging alternative to deep learning-based detection methods.

Nguyen et al. [12] investigated a deep learning-based approach for detecting steganographic content that may be exploited to conceal malicious information within digital images. The study employed SRNet (Steganalysis Residual Network), a CNN architecture specifically designed for image steganalysis. Unlike conventional CNN architectures that primarily learn semantic image content, SRNet incorporates steganalysis-specific residual processing to emphasize subtle noise and embedding artifacts introduced during information hiding. The model was evaluated using the ALASKA2 dataset, with appropriate strategies applied to address data imbalance and improve training stability. The experimental results demonstrated improved steganalysis performance compared with the reported baseline approaches. The study highlights the effectiveness of residual-based deep learning architectures for identifying subtle steganographic artifacts and strengthening image-based cybersecurity analysis.

Huo et al. [13] propose a Chaotic mapping-enHanced imAge Steganography nEtwork (CHASE), which pioneers to hide colour images in grey images and reduces the difference between the container image and the cover image through the image permutation method, so as to enhance the security of the steganography. The method demonstrates excellent steganalysis resistance in experiments and introduces Generative Adversarial Networks (GANs) to improve the image fidelity in large-capacity steganographic scales. The fusion of chaotic mapping and GAN optimisation enables the steganographic network to simultaneously balance security and image quality. The experimental results show that CHASE can keep the secret image with good invisibility under large-capacity steganographic scales, and at the same time, it can reveal the secret image with high fidelity, and its steganalysis-resistant capability is much better than other state-of-the-art methods.

Verma et al. [14] presents a tool in python to detect malware in widely used JPEG image format. The existing methods have mostly focused on finding steganography artifacts or used feature-based analysis that lacks revealing the hidden malign data. Unlike existing ones, the proposed tool alongside classification locates malicious content in JPEG images with revealing the found malign data along with its location as the output. This functionality to the best of our knowledge has not been found in the available literature. The tool has analyzed three types of JPEG images: malicious, benign, and stego images. Though malicious images are also stego ones, the paper refers to those hiding non-malicious data as the stego images. This is to evaluate the effectiveness of our tool in classifying the images with malicious and non-malicious data hidden. As a result, the tool has attained a low False Negative Rate (FNR) of 0.08 and False Positive Rate (FPR) of 0.001 with a better detection rate relative to state-of-the-art techniques.

2.1 Limitations of Existing Systems

- Existing systems often show limited capability in detecting low-payload malware hidden within steganographic images.
- CNN-based methods may capture local features

effectively but can have difficulty modeling long-range and global dependencies within images.

- Existing methods may be vulnerable to adversarial manipulation and advanced adaptive steganographic attacks.
- High-dimensional image processing can result in increased computational complexity and detection latency.
- Many deep-learning systems operate as black-box models, providing insufficient explanations for malware-detection decisions.

3. PROPOSED SYSTEM

In this section to described an the development of XViT-GAT: Explainable Vision Transformer–Graph Attention Network Framework for Robust Malware Detection in Steganographic Images, an advanced hybrid deep learning framework that integrates self-supervised representation learning, vision transformer-based feature extraction, graph attention-based artefact modelling, and explainable artificial intelligence, in order to address these limitations. Initially, the adaptive noise extraction module is implemented in the proposed framework to enhance the subtle steganographic residual patterns and suppress irrelevant image information. The hybrid Vision Transformer encoder effectively captures both local steganographic irregularities and long-range dependencies by processing the extracted representations through multi-head self-attention mechanisms. Additionally, complex relationships among spatial artefacts, frequency-domain variations, compression inconsistencies, and malware-related embedding characteristics are modelled using a Graph Attention Network (GAT) module. Figure 1 shows the architecture diagram for proposed method.

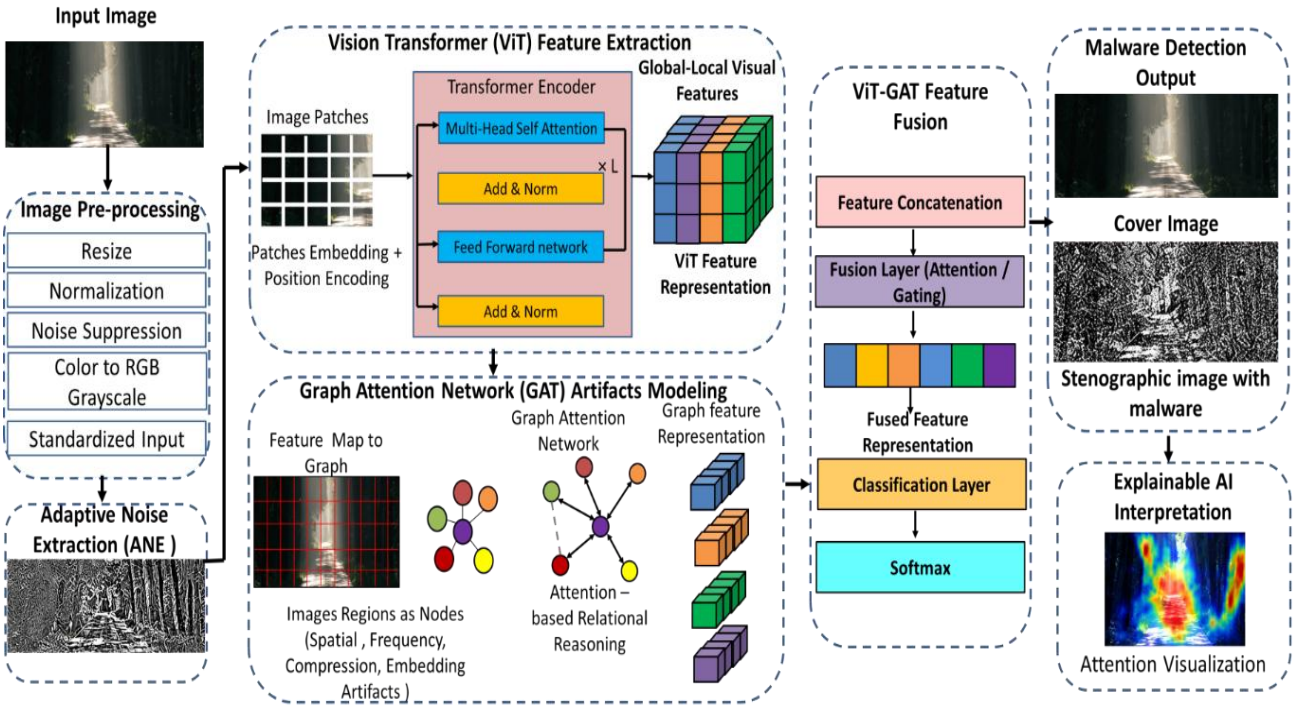


Fig 1: The architecture diagram for proposed method

3.1 Steganographic Image Acquisition

The initial stage involves acquiring cover, clean, and steganographic images for robust malware detection. Images are collected under diverse embedding algorithms, payload sizes, image formats, and compression conditions to represent realistic steganographic scenarios. Controlled stego samples containing malware-related payloads can additionally be generated and clearly separated from ordinary steganography samples. This diverse dataset provides the foundation for training and evaluating the proposed XViT-GAT framework.

3.2 Image Preprocessing and Standardization

The input images are preprocessed and standardized to provide a uniform representation while preserving the subtle pixel-level artifacts produced by steganographic embedding. Each image I is resized to a fixed spatial resolution $H \times W$, followed by pixel-intensity normalization to reduce variations caused by image size, intensity range, and acquisition conditions. Excessive filtering is avoided because hidden malware information may introduce very small statistical changes that are important for steganalysis. The resizing operation is expressed as

$$I_r = R(I, H, W) \tag{1}$$

Where $R(\cdot)$ represents the resizing function and I_r denotes the resized image. For an 8-bit image, pixel normalization is performed as

$$I_n(x, y, c) = \frac{I_r(x,y,c)}{255} \tag{2}$$

Where $I_n(x, y, c) \in [0,1]$is the normalized intensity at position (x, y)in channel c. To obtain a standardized representation, channel-wise standardization is further performed as

$$I_s(x, y, c) = \frac{I_n(x,y,c) - \mu_c}{\sigma_c + \epsilon} \tag{3}$$

Where μ_c and σ_c are the mean and standard deviation of channel c, respectively, and $\in$ is a small constant for numerical stability. The resulting standardized image I_s retains fine-grained steganographic traces while providing consistent input for

subsequent feature extraction and malware detection stages.

3.3 Adaptive Noise and Residual Extraction

Adaptive noise and residual extraction is applied to reveal weak steganographic artefacts that may be hidden within the dominant visual content of an image. The standardized image I_s is processed using an adaptive high-pass filter to suppress low-frequency structures, smooth regions, and general image content while retaining high-frequency pixel variations. The resulting residual map captures subtle embedding disturbances, abnormal noise patterns, and local pixel inconsistencies that may indicate the presence of concealed malware data. The basic residual extraction is formulated as

$$R(x,y) = I_s(x,y) - \hat{I}(x,y) \tag{4}$$

Where $I_s(x,y)$ represents the standardized input pixel and $\hat{I}(x,y)$ is the estimated local image content. Equivalently, high-pass residual extraction can be expressed as

$$R(x,y) = I_s(x,y) * K_{hp} \tag{5}$$

Where K_{hp} is the high-pass filtering kernel and $*$ denotes convolution. To emphasize locally abnormal residual responses, an adaptive weighting coefficient is defined as

$$\alpha(x,y) = \frac{\sigma_L(x,y)}{\sigma_L(x,y)+\lambda} \tag{6}$$

Where $\sigma_L(x,y)$represents the local standard deviation and λ is a small regularization parameter. The final adaptive residual map is obtained as

$$R_A(x,y) = \alpha(x,y)R(x,y) \tag{7}$$

Where $R_A(x,y)$ provides an enhanced representation of weak embedding artefacts and abnormal noise patterns for subsequent steganographic feature extraction.

3.4 Self-Supervised Representation Learning

Self-supervised representation learning is used to pretrain the feature encoder on large amounts of unlabeled image data so that it can learn robust and generalized representations without requiring manual annotations. For each residual-enhanced image R_A, two steganography-preserving augmented views are generated and passed through the same encoder f_θ. The encoder is trained to produce similar representations for different views of the same image while separating representations belonging to different images. This enables the model to capture subtle statistical, structural, and noise-related characteristics associated with steganographic embedding and improves robustness against previously unseen embedding patterns. The augmented views and their feature representations are expressed as

$$v_1 = T_1(R_A), \quad v_2 = T_2(R_A) \tag{8}$$

$$z_1 = f_\theta(v_1), \quad z_2 = f_\theta(v_2) \tag{9}$$

Where $T_1(\cdot)$and $T_2(\cdot)$ are augmentation functions, $f_\theta(\cdot)$is the trainable feature encoder, and z_1, z_2 are the corresponding latent representations. The similarity between two representations is calculated using cosine similarity as

$$sim(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\|_2 \|z_j\|_2} \tag{10}$$

The encoder is optimized using a contrastive self-supervised loss,

$$L_{SSL} = -log \frac{exp(sim(z_i,z_j)/\tau)}{\sum_{k\neq i} exp(sim(z_i,z_k)/\tau)} \tag{11}$$

Where (z_i, z_j)represents a positive pair obtained from the same image, $z_k$represents other samples, and τ is the temperature parameter. Minimizing L_{SSL} produces a generalized feature encoder that can subsequently be fine-tuned for robust detection of malware concealed within steganographic images.

3.5 Hybrid Vision Transformer Feature Extraction

The residual-enhanced representation R_A is supplied to the Vision Transformer (ViT) encoder to extract discriminative features associated with steganographic modifications. The input is divided into non-overlapping patches, and each patch is flattened and projected into a latent embedding space. Positional information is incorporated to preserve the spatial relationships among patches. Multi-head self-attention then models interactions between different image regions, allowing the network to capture both subtle local pixel abnormalities and long-range spatial dependencies that may indicate hidden malware payloads. The patch embedding is expressed as

$$z_i = x_i E + p_i \tag{12}$$

Where $x_i$denotes the flattened i-th residual patch, $E$is the learnable projection matrix, and $p_i$is the positional embedding. The complete input token sequence is formulated as

$$Z_0 = [z_{cls}; z_1; z_2; \dots; z_N] \tag{13}$$

Where z_{cls} is the classification token and N represents the total number of patches. For each attention head, the query, key, and value representations are calculated as

$$Q = ZW_Q, \quad K = ZW_K, \quad V = ZW_V \tag{14}$$

The self-attention operation is then defined as

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{15}$$

Where $d_k$denotes the dimensionality of the key vectors. Multiple attention heads are combined as

$$MHSA(Z) = Concat(head_1, head_2, \dots, head_h)W_o \tag{16}$$

Finally, the ViT feature representation is obtained as

$$F_{ViT} = ViT(R_A) \tag{17}$$

Where F_{ViT} contains integrated local and global steganographic features for subsequent graph-based relational learning and malware classification.

3.6 Multi-Domain Artefact Feature Construction

Multi-domain artefact feature construction is used to capture diverse traces generated by steganographic malware embedding

that may not be sufficiently represented in a single feature domain. Complementary spatial, residual, frequency-domain, compression, and embedding-related features are extracted from the standardized image and its residual representation. Spatial features characterize local intensity and texture variations, residual features emphasize weak embedding noise, frequency-domain features identify abnormal spectral changes, compression features capture inconsistencies in compressed image coefficients, and embedding-related features represent statistical deviations caused by hidden payloads. These complementary features are integrated into a unified representation to provide richer and more discriminative information for subsequent graph-based malware detection. The domain-specific features are expressed as

$$f_{sp} = \phi_{sp}(I_s), \quad f_{res} = \phi_{res}(R_A) \tag{18}$$

$$f_{freq} = \phi_{freq}\big(\mathrm{T}(I_s)\big), \quad f_{comp} = \phi_{comp}(I_s) \tag{19}$$

$$f_{emb} = \phi_{emb}(I_s, R_A) \tag{20}$$

Where $\phi(\cdot)$denotes the corresponding feature extraction function and T(·)represents a frequency transformation such as DCT or DWT. The extracted features are concatenated to form the multi-domain representation,

$$F_{MD} = f_{sp} \| f_{res} \| \| f_{freq} \| \| f_{comp} \| f_{emb} \tag{21}$$

Where || denotes concatenation. The combined features are then projected into a common latent feature space as

$$F_{MDA} = \sigma(W_{MD} F_{MD} + b_{MD}) \tag{22}$$

Where W_{MD} and b_{MD} are learnable weight and bias parameters, respectively, and $\sigma(\cdot)$represents a nonlinear activation function. The resulting F_{MDA} provides a comprehensive representation of heterogeneous steganographic artefacts for subsequent GAT-based relational feature learning and malware classification.

3.7 Graph Construction and GAT-Based Reasoning

Graph construction and GAT-based reasoning are employed to capture structural relationships among significant image regions and steganographic artefacts. The feature tokens obtained from the ViT and multi-domain feature construction stages are represented as graph nodes, while edges connect nodes exhibiting strong spatial or feature-level relationships. Each node i is associated with a feature vector h_i, and cosine similarity can be used to determine the relationship between nodes i and j as

$$s_{ij} = \frac{h_i^T h_j}{\|h_i\|_2 \|h_j\|_2} \tag{23}$$

An edge is established when the similarity exceeds a predefined threshold δ,

$$(i,j) \in E \quad if \;\; s_{ij} \geq \delta \tag{24}$$

The GAT subsequently calculates an attention score for each connected pair of nodes as

$$e_{ij} = LeakyReLU\big(a^T [Wh_i || Wh_j]\big) \tag{25}$$

Where W is the learnable weight matrix, a is the attention vector, and ‖ represents concatenation. The attention coefficients are normalized using the softmax function,

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k \epsilon N(i)} exp(e_{ik})} \tag{26}$$

Finally, the graph-aware representation of node i is obtained by aggregating its neighbouring features according to their learned importance,

$$h_i' = \sigma\big(\sum_{j \epsilon N(i)} \alpha_{ij} W h_j\big) \tag{27}$$

Where $N(i)$ denotes the neighbouring nodes and $\sigma(\cdot)$is a nonlinear activation function. This process enables the GAT to emphasize informative artefact relationships and generate structurally enriched features for robust steganographic malware detection.

3.8 ViT–GAT Feature Fusion and Malware Classification

The ViT–GAT feature fusion stage combines the global contextual information extracted by the Vision Transformer with the structural and relational artefact information learned by the Graph Attention Network. The ViT features F_{ViT} characterize long-range spatial dependencies and distributed embedding abnormalities, whereas the GAT features F_{GAT} represent relationships among suspicious image regions. These complementary representations are concatenated and projected into a unified feature space to generate a discriminative representation for malware detection. The fusion operation is defined as

$$F_{cat} = F_{ViT} \parallel F_{GAT} \tag{28}$$

Where ‖denotes feature concatenation. The fused representation is obtained as

$$F_{fusion} = \sigma\big(W_f F_{cat} + b_f\big) \tag{29}$$

Where W_f and b_f are learnable fusion parameters and $\sigma(\cdot)$denotes a nonlinear activation function. For binary detection, the probability that an image contains steganographically embedded malware is calculated using

$$\hat{y} = \frac{1}{1 + exp[-W_c F_{fusion} + b_c]} \tag{30}$$

The final classification is determined as

$$\hat{C} \begin{cases} \frac{Clean}{Cover}, & \hat{y} < \tau_c \\ Malware - Containing\ Stegnographic\ Image, & \hat{y} \geq \tau_c \end{cases} \tag{31}$$

Where τ_c is a decision threshold selected using validation data. When multiple malware or embedding categories are available, multiclass probabilities can be computed using

$$P(y = k) = \frac{exp(o_k)}{\sum_{j=1}^{K} exp(o_j)}, \quad \hat{C} = \underset{k}{argmax}\, P(y = k) \tag{32}$$

Where K represents the number of malware or embedding classes and o_k is the classification logit corresponding to class k.

3.9 Explainable AI-Based Interpretation

Explainable AI (XAI)-based interpretation is incorporated into the XViT-GAT framework to provide transparent evidence for each malware detection decision. Attention visualization can highlight influential ViT patches and GAT node relationships, while SHAP, Integrated Gradients (IG), or a Grad-CAM-

compatible attribution method can quantify the contribution of image regions and learned features to the predicted class. These attribution maps help identify whether the model's decision is driven by meaningful residual disturbances, frequency abnormalities, or suspicious embedding-related regions rather than irrelevant visual content. For Integrated Gradients, the contribution of input feature $x_i$to the model output $F(X)$ is defined as

$$IG_i(X) = (x_i - x_i') \int_0^1 \frac{\partial F\left(x' + \alpha(x - x')\right)}{\partial x_i} d_\alpha \quad (33)$$

Where x is the input representation, $x'$is a baseline input, and α controls the interpolation between the baseline and actual input. For SHAP-based interpretation, the prediction can be represented as

$$F(X) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i \quad (34)$$

Where $\emptyset_0$ denotes the baseline prediction and $\emptyset_i$represents the contribution of the i-th feature to the detection result. A Grad-CAM-compatible localization map can be formulated as

$$L_{XAI}^c = ReLU(\sum_k \alpha_k^c A^k) \quad (35)$$

With

$$\alpha_k^c = \frac{1}{Z} \sum_{u,v} \frac{\partial y^c}{\partial A_{uv}^k} \quad (36)$$

Where A^k represents the k-th feature map, $y^c$is the score for class c, and α_k^cmeasures the importance of that feature map. The resulting attribution evidence highlights the regions and features that most strongly influence the clean/cover versus malware-containing steganographic image prediction, improving the interpretability and forensic usefulness of the proposed framework.

4. RESULT AND DISCUSSION

The experiments for Malware Detection in Steganographic Images were conducted using Python with deep-learning libraries on a workstation equipped with an Intel Core i9 processor, 32 GB RAM, and NVIDIA RTX 3090 GPU (24 GB). The steganographic image dataset was divided into training, validation, and testing sets. All images were resized and normalized before training, and the proposed model was evaluated using accuracy, precision, recall, F1-score, and AUC to assess its malware-detection performance. In this research using the existing systems are SiaIRNet [15], GBRAS-Net [16], SRNet [17], EfficientNet [18].

4.1 Dataset

The Break Our Steganographic System (BOSS) dataset [19] is a widely used benchmark for evaluating digital image steganalysis and steganography detection methods. The BOSSBase collection contains 10,000 grayscale images with a resolution of 512 × 512 pixels, acquired using multiple digital cameras and subsequently standardized for experimental use. The original images serve as cover images, while corresponding stego images can be generated by embedding hidden payloads using algorithms such as HUGO, WOW, S-UNIWARD, and MiPOD at different embedding rates. For the proposed malware detection framework, BOSS images can be used to construct controlled cover/stego pairs by embedding benign or appropriately prepared malware-derived payload representations. This enables evaluation of the model's ability to identify subtle embedding artifacts and distinguish clean cover images from steganographic images containing concealed payloads. Figure 2 shows the sample images of dataset.



Fig 2: BOSS Dataset

Table 1. Accuracy Analysis for Proposed Method

Number of Data	Dataset									
	Cover					Stego				
	SiaIRNet	GBRAS-Net	SRNet	EfficientNet	Proposed	SiaIRNet	GBRAS-Net	SRNet	Efficient Net	Proposed
100	82.36	84.71	86.24	87.63	91.42	80.18	73.27	85.11	86.45	90.28
200	85.74	87.82	89.16	90.48	93.76	84.26	76.38	88.07	89.21	92.64
300	88.63	70.14	81.72	82.86	95.38	87.42	79.05	80.36	81.58	94.47
400	70.81	82.27	73.65	84.53	97.16	89.73	81.18	72.44	83.62	96.28
500	82.46	73.85	75.12	87.07	98.72	81.38	82.76	74.18	85.21	98.86

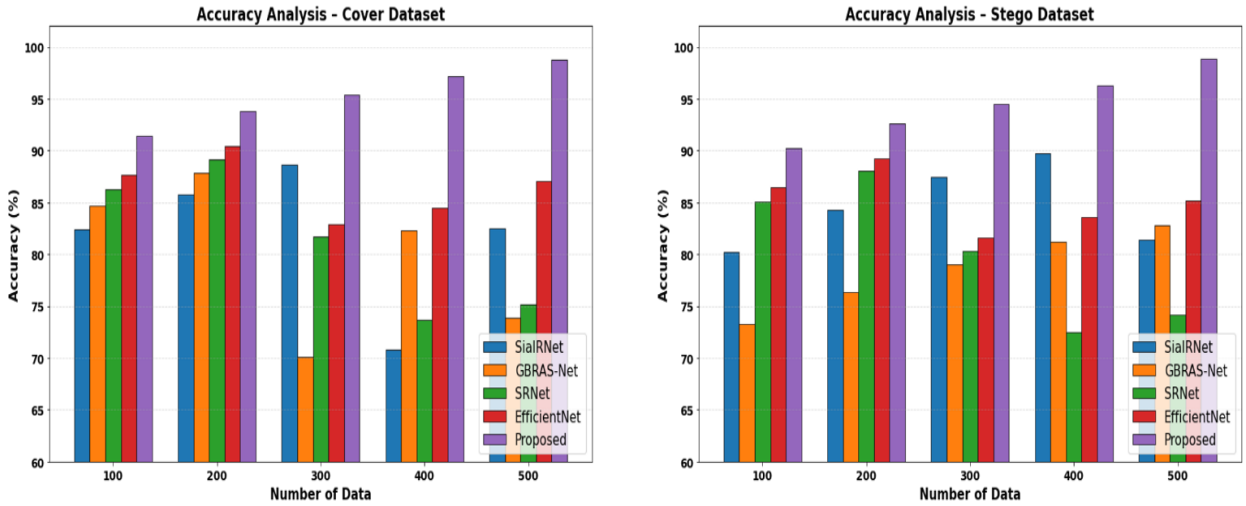


Figure 3: Accuracy Analysis for Proposed Method

Table 1 and Figure 3 presents the accuracy analysis of the proposed method compared with SialRNet, GBRAS-Net, SRNet, and EfficientNet for both cover and stego images across different data sizes. The proposed method consistently achieves higher accuracy than the comparative approaches. For cover images, its accuracy increases from 91.42% at 100 samples to 98.72% at 500 samples, while for stego images, it improves from 90.28% to 98.86%. Although the existing methods exhibit considerable fluctuations as the number of samples increases, the proposed approach demonstrates a clear improvement in performance and maintains comparatively high accuracy across all experimental settings. In particular, at 500 samples, the proposed method achieves the highest accuracies of 98.72% for cover images and 98.86% for stego images, indicating its effectiveness in learning discriminative features for distinguishing cover and steganographic images

5. CONCLUSION AND FUTURE WORK

This study presents XViT-GAT, an explainable hybrid framework for robust malware detection in steganographic images. The framework integrates Vision Transformer (ViT)-based global contextual learning with Graph Attention Network (GAT)-based relational reasoning to capture subtle embedding artefacts and complex dependencies among image regions. Multi-domain features derived from spatial, residual, frequency, compression, and embedding characteristics further strengthen the representation of steganographic traces. In addition, Explainable AI (XAI) techniques provide interpretable evidence about the image regions and features contributing to detection decisions. The framework is designed to improve the discrimination of clean/cover and malicious steganographic images while enhancing robustness and interpretability for cybersecurity and digital-forensic applications. Future work can investigate cross-dataset validation, unseen steganographic algorithms, adversarial robustness, lightweight deployment, and real-world malware-bearing steganographic samples to further establish the framework's practical effectiveness.

6. REFERENCES

[1] Liguori, A., Zuppelli, M., Gallo, D., Guarascio, M., and Caviglione, L. 2026. A deep learning-based approach for stegomalware sanitization in digital images. Journal of Intelligent Information Systems. 64, 1, 121-144.

[2] De La Croix, N. J., Ahmad, T., and Han, F. 2024. Comprehensive survey on image steganalysis using deep learning. Array. 22, 100353.

[3] Roshan, N., Barik, D., and Roseline, S. A. May 2024. MalViT: An Approach to Enhancing Malware Detection. In 2024 International Conference on Advances in Computing. Communication and Applied Informatics (ACCAI) IEEE. 1-8.

[4] Belal, M. M., and Sundaram, D. M. 2024. Multi-variants vision transformer-based malware image classification model using multi-criteria decision-making. Journal of Intelligent & Fuzzy Systems. 46, 5-6, 11331-11351.

[5] Cui, B., Wang, H., Qi, Y., Chen, H., Yuan, Q., Liu, D., and Zhou, X. 2025. HERL-ViT: A hybrid enhanced vision transformer based on regional-local attention for malware detection. Computers, Materials & Continua. 85, 3, 5531-5553.

[6] Chen, S., Lang, B., Liu, H., Chen, Y., and Song, Y. 2024. Android malware detection method based on graph attention networks and deep fusion of multimodal features. Expert systems with applications. 237, 121617.

[7] Tanveer, M. U., Munir, K., Alyamani, H. J., Hassan, S. R., Sheraz, M., and Chuah, T. C. 2025. Graph-augmented multi-modal learning framework for robust ndroid malware detection. Scientific Reports. 15, 1, 38341.

[8] Baghirov, E. 2025. A comprehensive investigation into robust malware detection with explainable AI. Cyber Security and Applications. 3, 100072.

[9] Ipek, M. C., and Sen, S. 2026. Explainable android malware detection and malicious code localization using graph attention. Journal of Information Security and Applications. 98, 104385.

[10] Kuznetsov, O., Frontoni, E., Chernov, K., Kuznetsova, K., Shevchuk, R., and Karpinski, M. 2024. Enhancing steganography detection with AI: fine-tuning a deep residual network for spread spectrum image steganography. Sensors. 24, 23, 7815.

[11] Płachta, M., Krzemień, M., Szczypiorski, K., and Janicki, A. 2022. Detection of image steganography using deep learning and ensemble classifiers. Electronics. 11, 10, 1565.

[12] Nguyen, Q. H., Nguyen, D. M., Bui, H. V. A., Huynh, K. Q., Nguyen, H. C. T., and Bui, N. D. T. November 2025. Malware Detection in Images Using AI: A Deep Learning

Approach for Steganography Detection. In International Conference on Data-Processing and Networking. Cham: Springer Nature Switzerland. 106-118.

[13] Huo, L., Chen, R., Wei, J., and Huang, L. 2024. A high-capacity and high-security image steganography network based on chaotic mapping and generative adversarial networks. Applied Sciences. 14, 3, 1225.

[14] Verma, V., Muttoo, S. K., and Singh, V. B. 2022. Detecting stegomalware: malicious image steganography and its intrusion in windows. In Security, privacy and data analytics: Select proceedings of ISPDA 2021 Singapore: Springer Singapore. 103-116.

[15] Li, H., Wang, J., Xiong, N., Zhang, Y., Vasilakos, A. V., and Luo, X. 2023. A siamese inverted residuals network image steganalysis scheme based on deep learning. ACM Transactions on Multimedia Computing, Communications and Applications. 19, 6, 1-23.

[16] Reinel, T. S., Brayan, A. A. H., Alejandro, B. O. M., Alejandro, M. R., Daniel, A. G., Alejandro, A. G. J., and Raul, R. P. 2021. GBRAS-Net: a convolutional neural network architecture for spatial image steganalysis. IEEe Access, 9, 14340-14350.

[17] Nguyen, Q. H., Nguyen, D. M., Bui, H. V. A., Huynh, K. Q., Nguyen, H. C. T., and Bui, N. D. T. November 2025. Malware Detection in Images Using AI: A Deep Learning Approach for Steganography Detection. In International Conference on Data-Processing and Networking Cham: Springer Nature Switzerland. 106-118.

[18] Judy, S., and Khilar, R. July 2025. Bat-Optimized Deep Learning Framework for Steganographic Malware Detection in Jpegs. In 2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI) IEEE. 547-554.

[19] https://www.kaggle.com/datasets/lijiyu/bossbase