Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

MULTIMEDIA STEGANALYSIS USING HYBRID CNN AND TRANSFORMER

Aroob Mukhtar[*1], Farhan Hassan[2], M. Madni[3], Umar Daraz[4]

[*1,2,3,4]*Department of Information and Communication Engineering, the Islamia University of Bahawalpur, Bahawalpur, Pakistan*

[*1]aroobmukhtar850@gmail.com

DOI: https://doi.org/10.5281/zenodo.20197136

Keywords
Multimedia Steganalysis, Hybrid CNN–Transformer, Deep Learning, Cybersecurity, Image Steganalysis, Audio Steganalysis, Video Steganalysis, Transformer Networks, Multimedia Forensics, Hidden Information Detection

Article History
Received: 16 March 2026
Accepted: 25 April 2026
Published: 13 May 2026

Corresponding Author: *
Aroob Mukhtar

Abstract
Steganography enables covert communication by concealing secret information within digital multimedia content such as images, audio, and video files. The increasing misuse of steganographic techniques in cybercrime and covert communication underscores the urgent need for effective multimedia steganalysis systems. This study introduces a unified multimedia steganalysis framework utilizing a Hybrid CNN–Transformer architecture to detect hidden information across diverse multimedia modalities. The framework integrates the local feature extraction strengths of Convolutional Neural Networks (CNNs) with the global contextual learning capabilities of Transformer encoders to identify spatial, spectral, and temporal steganographic artifacts. Publicly available datasets, such as BOWSBASE, BOWS2, TIMIT, ESC-50, LibriSpeech, HMDB51, UCF-101, and Kinetics-400, are employed for experimental evaluation. The model is assessed using various embedding techniques and multimodal late fusion for final classification. Results indicate that the proposed framework outperforms standalone CNN and Transformer models, achieving an overall accuracy of 96.4%, and demonstrating enhanced robustness and generalization across multimedia modalities.

I. INTRODUCTION

Digital communication has transformed modern society by facilitating large-scale multimedia data exchange across interconnected systems. This expansion has heightened the demand for secure communication mechanisms to ensure data confidentiality, integrity, and authenticity [1]. Although various security techniques exist, ongoing advancements are necessary to address emerging threats and evolving attack strategies.

Steganography is a widely used technique for covert communication in which secret information is embedded within digital media such as images, audio, video, and text. While cryptography focuses on securing the data content of a message, steganography hides the very fact that there is a message, making it hard to detect [2], [3]. The possible malicious usage of the technique, including its employment for covert communication and data leakage, makes the development of steganalysis techniques crucial nowadays.

Steganalysis is the process of detecting hidden information within digital media, typically formulated as a binary classification problem distinguishing between cover and stego data [3]. Early steganalysis methods relied on handcrafted feature extraction techniques, such as Spatial Rich Models (SRM) and related statistical approaches, combined with machine learning classifiers, including ensemble classifiers and Support Vector

Machines (SVMs) [4], [5]. Although these approaches performed well for classical steganographic techniques, they require domain expertise and often fail to generalize modern adaptive embedding methods [1], [6].

Recent advancements in deep learning have significantly improved steganalysis performance by providing the capability to automatically learn features from the given payload-free raw data. CNN remains a popular choice to effectively model local spatial dependencies and noise residual patterns that could be introduced by the embedding process [7], [8]. Further advancement on the CNN model that improves the detection performance has been seen in other architectures such as Xu-Net [9], as well as in deeper residual networks, such as SRNet [10]. These deep learning models have been reported to outperform feature-based steganalysis at both detection accuracy and AUC scores, on different media types [1].

More recently, Transformer-based models have demonstrated the ability to effectively model long-range dependencies in the payload-free raw data through the self-attention mechanism. The recently proposed Transformer models [11], as well as their extension from NLP to vision tasks through vision Transformer (ViT) [12], allow global context modeling and naturally fit in multimedia data for detection tasks.

Hierarchically using the Transformer architecture, the Swin Transformer can efficiently perform large-scale global modeling [13]. While CNNs work well on extracting local features, the Transformers naturally perform best in capturing long-range relationships requiring large amounts of data. Consequently, hybrid CNN–Transformer architectures have gained significant attention in multimedia steganalysis research. Such models benefit from the local detection of features while leveraging the global description of the data to perform higher detection accuracy [14], [15].

Several research challenges remain, such as generalization of detection accuracy to different datasets, and also cross-media sensitivity due to the data distribution change. Cover-source mismatch has been established to be the main contributor to the failure of real-world steganalysis detection [16].

In this paper, we focus on multimedia steganalysis utilizing hybrid CNN-Transformers models and explore their potential in different media types.

II. Literature Review

A. Traditional steganalysis approaches:

Conventional steganalysis approaches are based on handcrafted feature extraction combined with statistical modeling, such as Spatial Rich Models (SRM), which extract high-dimensional noise residual features to describe embedding artifacts [4], generally fed into an ensemble classifier for classification [5]. Several variants of feature-based approaches have been proposed to improve detection performance on different stego schemes [17]. These approaches achieve strong performance when carefully designed results for early techniques, but manually designing features limits the flexibility of their features, leading to reduced adaptability to modern embedding techniques and degrading performance on unseen data or various data distributions [6], [18].

B. CNN based steganalysis approaches:

Deep learning-based methods have attracted significant attention and have outperformed traditional approaches by learning features automatically and efficiently. Qian et al. [7] confirmed the feasibility of CNNs by learning independent discriminative features from the image data. Afterwards, some optimized CNN architectures, including Yedroudj-Net [8], Xu-Net [9], and SRNet[10], which extracted features to analyze the spatial disturbances in natural images from the subtle differences between the cover and stego images, achieved good results in image steganalysis. Cascading CNNs are applied in audio steganalysis as well [19], in video steganalysis, extracting temporal and spectral characteristics of signals [20]. Valuable information resulted from the CNN.

Li, Zhonghao, and colleagues present a steganography technique for HEVC videos based on a PU approach utilizing Wide Residual-Net (PWRN), which achieves high embedding efficiency while maintaining visual quality. This method leverages a super-resolution CNN equipped with wide residual filters to reconstruct

I-pictures, thereby enhancing P-picture prediction, reducing bitrate, and resisting PU-targeted steganalysis [20]. Reinel, Tabares-Soto, and others propose a novel CNN for spatial image steganalysis, incorporating a preprocessing phase with filter banks, depthwise and separable convolutions, as well as skip connections to more effectively accentuate steganographic noise [21].
However, the effective modeling capacity of CNNs is limited when capturing long-range dependencies due to their localized receptive fields.

C. Transformer-based approaches:
With the evolution of deep Transformer architectures like Vision Transformer (ViT) [12], the ability of modeling long-range dependencies in images has been greatly enhanced, which is applicable to analyzing the distribution of the embedding distribution spatial relationship in multimedia steganalysis. Substantial improvements have been achieved in modeling global contextual information for vision tasks using hierarchical Transformer architectures such as SWin Transformer [13], which can learn performances of global long-range dependencies.
C. Zhang and their team introduce TENet, a steganalysis framework based on Transformers designed for VoIP, which employs codeword and position embeddings to reveal concealed representations of VoIP streams aimed at detecting QIM-based steganography [22].
However, Transformer models have heavy numbers of parameters, require large training sets, are computationally intensive, which limits their application to many datasets in steganalysis, but have limitations on modeling local details when used independently.

D. Hybrid approaches:
Due to the limitations of CNNs and the global sensitivity of the Transformer architecture. Hybrid CNN-Transformer models have been proposed. For instance, Luo et al. [14] designed a convolutional vision transformer framework for image steganalysis, which combined the ability of local features learning of CNNs with global context modeling of Transformer, achieved high detection accuracy.
Bravo-Ortiz et al. [15] developed hybrid architectures combining the convolutional network with the attention module. The fusion of local features learning with global context modeling helps in learning more effective representations. Wei et al. [23] proposed CTNet, a hybrid CNN–Transformer model for color image steganalysis that combines residual-based preprocessing with local and global feature extraction. Their results show improved detection performance over existing methods.
Wang et al. [24] introduce a hybrid model called CTS-Net (CNN-Transformer image steganography network) designed for image steganalysis. This model effectively captures dependencies in both local and global features of steganographic signals.
A hybrid of CNN and Transformer, proposed by Peng et al., was designed for audio, which involves the extraction of features using multi-scale CNN and Transformer components [25].

III. Methodology
The proposed system uses a Hybrid CNN and a Transformer to detect steganalysis in multimedia formats such as audio, video, and images.
For audio and video datasets that do not already contain embedded steganography, steganographic data can be added to prepare them for training. For image steganalysis, sample data should be taken from the BOWSBASE-1.01 and BOWS2 dataset, which already includes steganography.
The detection model integrates Convolutional Neural Networks (CNNs) and Transformers. CNN layers extract local spatial and temporal artifacts of steganography, while the Transformer component identifies global dependencies. The features are then used for the classification of whether there is hidden information within the media.
The complete methodology consists of the following major stages:

1. Multimedia dataset collection
2. Controlled steganographic embedding generation
3. Modality-specific preprocessing
4. CNN-Based Feature Extraction
5. Transformer-Based Contextual Modeling
6. Modality-Wise Classification

7. Multimodal late fusion and final decision generation

This pipeline ensures consistent detection performance across multiple multimedia modalities

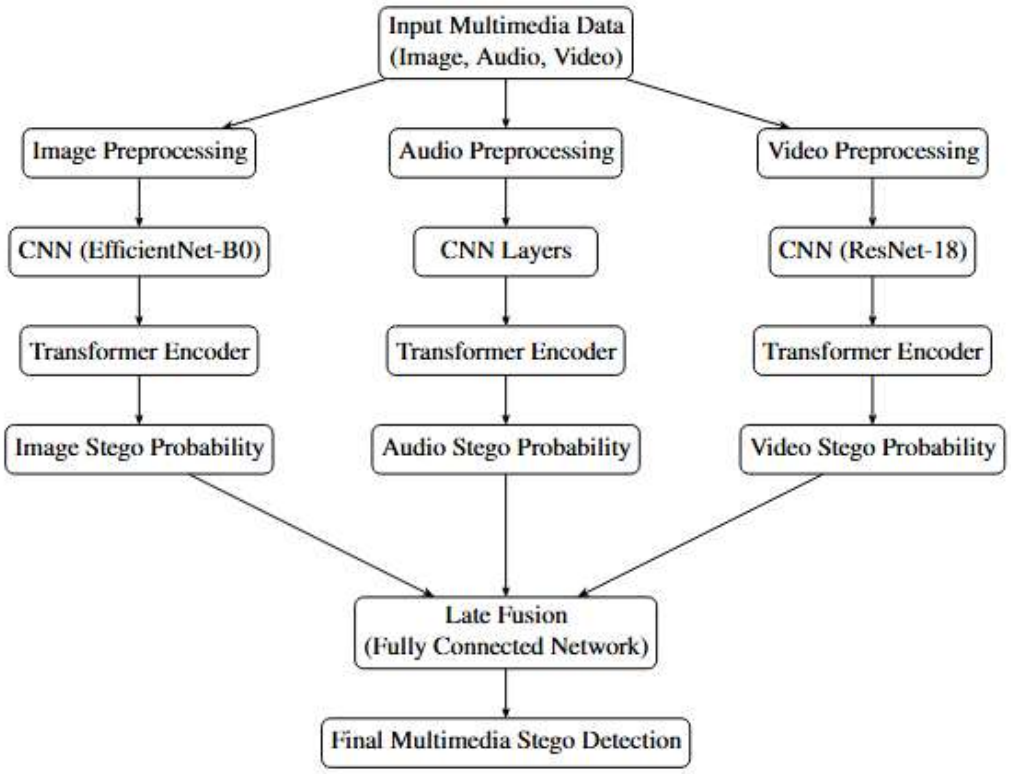

Figure 1. Overall architecture of the proposed Hybrid CNN–Transformer multimedia steganalysis framework.

3.1 Multimedia dataset collection:

Different publicly available multimedia datasets are collected to support the development and evaluation of the proposed multimedia steganalysis framework. These datasets ensure variety in content, format, and signal characteristics in images, audio, and video modalities. This diversity in the datasets improves the generalization of the proposed detection models.

3.1.1 Image Dataset Collection:

The BOWSBASE+BOWS2 dataset is used for image steganalysis. This dataset contains both BOWSBASE samples as well as BOWS2 samples with different payloads and different stego algorithms. An equal number of stego or cover images is used for the training modality. These characteristics make this dataset diverse for training a hybrid image modality for a multimedia steganalysis system. The combined dataset contains grayscale images along with embedded steganography generated using well-known algorithms. Hence, the images already include steganographic embeddings; the dataset is used in its original form without further modification.

3.1.2 Audio Dataset Collection:

Audio data is collected from:

- TIMIT Corpus
- ESC-50
- LibriSpeech

Approximately 9,988 audio samples are extracted from LibriSpeech to increase dataset diversity.

The datasets provide:

- Clean speech recordings

- Environmental sounds
- Diverse acoustic conditions
- Multiple speaker variations

Since these datasets initially contain clean audio signals, steganographic embeddings are later applied using different audio hiding techniques.

3.1.3 Video Dataset Collection:

For video steganalysis, the following datasets are utilized:

- HMDB51
- UCF-101
- Kinetics-400

Approximately 16,000 video samples are collected from UCF-101 and Kinetics-400 datasets.

The datasets are selected because they contain:

- Rich temporal information
- Human activity variations
- Complex motion patterns
- Real-world visual scenes

The collected videos are initially clean and later converted into stego-video samples using different embedding methods.

For multimodal fusion experiments, video frames and audio streams are extracted from UCF-101 videos.

3.2 Controlled steganographic embedding generation:

Steganography embedding is a technique used to add hidden information to a clean multimedia dataset in a controlled way. This step creates realistic stego samples by using different embedding techniques in the spatial, temporal, and frequency domains. The main aim of this step is to detect hidden data or embeddings without changing media quality.

3.2.1 Image Steganography Embedding:

The BOSSBASE+BOWS2 dataset already contains steganographic samples generated using:

- HILL
- S-UNIWARD
- WOW
- MiPOD
- HUGO

3.2.2 Audio Steganography Embedding:

The process of audio steganography embedding is performed on the clean audio signal to bring about the unnoticeable changes that will hold the concealed information.

Techniques include:

- LSB embedding
- Echo hiding
- Phase modulation
- Spread spectrum
- Backmasking

3.2.3 Video Steganography Embedding:

Video steganography embedding hides information by using the spatial, temporal, and motion features found in video sequences.

Techniques include:

- LSB Embedding
- Frequency-Domain Embedding
- GAN-Style Residual Embedding
- DCT-Domain Embedding
- Compression-Aware Embedding
- Hybrid Embedding
- Motion-Sensitive Embedding

3.3 Modality-specific preprocessing:

Preprocessing plays a vital role in the multimedia steganalysis framework by preparing various types of images, audio, and video data for hybrid CNN and Transformer models.

Each data type, images, audio, and video, has its own preprocessing steps to handle its unique structure.

3.3.1 Image Preprocessing:

Image preprocessing is performed using PyTorch transformation pipelines.

The preprocessing stages include:

1. Image resizing to 224 ×224 pixels
2. Tensor conversion
3. Pixel normalization to the range [0,1]

The resizing operation ensures compatibility with EfficientNet-B0.

3.3.2 Audio Preprocessing:

Raw audio signals are converted into Mel-spectrogram representations.

The preprocessing stages include:
1. Audio loading
2. Resampling to 16 kHz
3. Mel-spectrogram extraction using 64 Mel bands
4. Logarithmic scaling
5. Padding or truncation to 256 frames

This representation preserves hidden spectral information for CNN and Transformer processing.

3.3.3 Video Preprocessing:

Video preprocessing focuses on generating fixed-length frame sequences.

The preprocessing stages include:
1. Video decoding using OpenCV
2. Extraction of 16 frames per video
3. Frame resizing to 224 ×224 pixels
4. Tensor conversion
5. Pixel normalization

Videos containing fewer than 16 frames are padded using the final frame.

3.4 CNN-Based Feature Extraction:

The convolution operation performed by CNN layers can be represented as:

$$F_{i,j}^{(k)} = \sigma\left(\Sigma_{m,n} W_{m,n}^{(k)} X_{i+m,j+n} + b^{(k)}\right)$$

where:
- W represents convolution kernels
- X represents input feature maps
- b represents bias terms
- σ denotes activation function
- F represents extracted feature maps

CNN layers learn local embedding artifacts and residual noise patterns.

3.5 Transformer-Based Contextual Modeling:

Transformer encoders utilize self-attention mechanisms to model global dependencies.

The self-attention operation is defined as:

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

where:
- Q denotes query vectors
- K denotes key vectors
- V denotes value vectors
- represents key dimensionality

The Transformer encoder captures long-range contextual relationships across multimedia features.

3.6 Modality Wise Classification:

3.6.1 Image Steganalysis Module:

The image steganalysis branch uses EfficientNet-B0 as the CNN backbone.

- CNN Backbone:

EfficientNet-B0 extracts local spatial features and residual artifacts introduced by steganographic embedding.

- Transformer Encoder:

CNN feature maps are transformed into sequential embeddings and processed through Transformer encoders.

- Classification Layer:

A fully connected binary classifier predicts whether the image is stego or cover.

3.6.2 Audio Steganalysis Module:

The audio branch processes Mel-spectrograms using CNN and Transformer components.

- CNN Layers:

Convolutional layers extract local spectral distortions.

- Transformer Encoder:

Transformer encoders model long-range temporal relationships.

- Classification:

The extracted embeddings are passed into fully connected layers for binary classification.

3.6.3 Video Steganalysis Module:

The video branch analyzes both spatial and temporal relationships.

- CNN Encoder:

ResNet-18 extracts frame-level spatial features.

- Transformer Encoder:

Transformer encoders model temporal dependencies between frames.

- Classification:

Temporal embeddings are classified into stego and non-stego categories.

3.7 Multimodal late fusion and final decision generation:

The late fusion module combines feature representations extracted from image, audio, and video branches.

- Fusion Strategy:

Feature vectors from all modalities are concatenated and passed into a lightweight fully connected network.

- Final Prediction:

The fused representation generates the final multimedia steganalysis decision.

This multimodal fusion improves robustness and detection performance.

3.8 Experimental Setup and Evaluation Metrics:

3.8.1 Model Training Configuration:

- **Framework:** PyTorch
- **Optimizer:** Adam
- **Learning Rate:** 0.0001
- **Batch Size:** 16
- **Epochs:** 16
- **Loss Function:** Binary Cross Entropy (BCE)

3.8.2 Loss Function:

Binary Cross Entropy (BCE) loss is used for optimization.

The BCE loss function is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})]$$

Where:

- y_i = actual label
- $\hat{y_i}$ = predicted probability
- N = total number of samples
- L = Binary Cross Entropy loss

This is the standard BCE loss equation used in deep learning papers.

3.8.3 Dataset Split Strategy:

The datasets are divided using stratified splitting to maintain class balance.

The dataset split configuration is:

- 70% Training
- 15% Validation
- 15% Testing

This strategy ensures reliable model evaluation and prevents overfitting.

3.8.4 Evaluation Metrics:

The proposed framework is evaluated using Accuracy, Precision, Recall, and F1-score.

- Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

- Precision:

$$\Pr ecision = \frac{TP}{TP+FP}$$

- Recall:

$$Recall = \frac{TP}{TP+FN}$$

- F1-Score:

$$F1 = 2 \times \frac{\Pr ecision \times Recall}{\Pr ecision + Recall}$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

IV. Result and Discussion

The proposed hybrid CNN–Transformer framework was evaluated on image, audio, and video steganalysis tasks using the datasets and preprocessing methodology outlined in Section III. The experiments assessed the architecture's ability to detect hidden information across multiple multimedia modalities while maintaining balanced performance and generalization.

Model training was performed using the PyTorch framework with the Adam optimizer, a learning

rate of 0.0001, a batch size of 16, and 16 epochs. The datasets were partitioned into training, validation, and testing sets using a 70:15:15 stratified split to ensure class balance and mitigate biased learning.

4.1 Image Steganalysis Results:

The image steganalysis branch was evaluated using the BOWSBASE+BOWS2 dataset containing both cover and stego images generated using HILL, WOW, MiPOD, S-UNIWARD, and HUGO embedding algorithms.

The hybrid EfficientNet-B0 and Transformer architecture demonstrated strong detection capability for hidden image embeddings. CNN layers effectively extracted local residual artifacts, while the Transformer encoder captured long-range spatial dependencies that improved classification consistency.

Model	Accuracy	Precision	Recall	F1-Score
CNN Only	91.8%	91.1%	92.4%	91.7%
Transformer Only	90.6%	89.8%	91.2%	90.5%
Hybrid CNN-Transformer	95.2%	94.8%	95.7%	95.2%

Table 1. Image Steganalysis Results

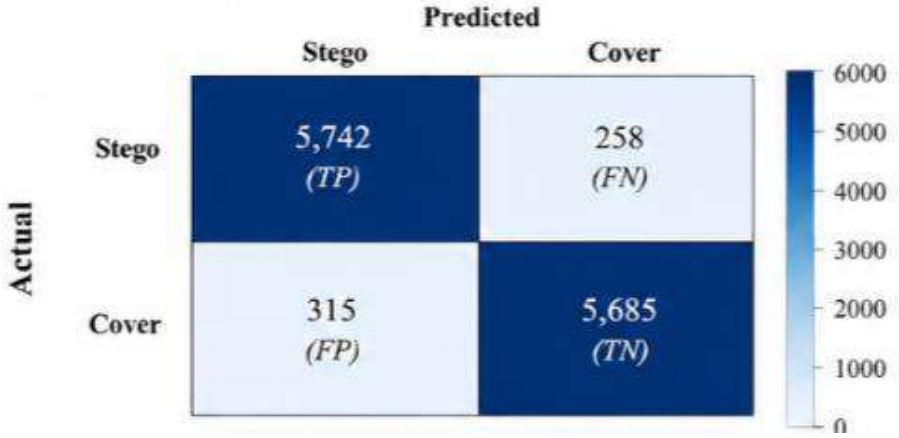

Figure 2. Confusion matrix for image steganalysis.

The results in Figure 2 and Table 1 showed that the hybrid model outperformed both the CNN-only and the Transformer-only architecture.

The CNN-only model successfully detected local embedding noise patterns but showed limitations in capturing broader contextual relationships. Conversely, the Transformer-only model captured global dependencies but was less effective in identifying fine-grained embedding artifacts.

The hybrid architecture combined both advantages, leading to improved detection accuracy and balanced precision-recall performance.

4.2 Audio Steganalysis Results:

The audio steganalysis module was evaluated with audio samples from the TIMIT, ESC-50, and LibriSpeech datasets. This evaluation followed controlled steganographic embedding using LSB embedding, echo hiding, phase modulation, spread spectrum, and backmasking techniques.

Mel-spectrogram preprocessing helped the CNN layers learn the spectral distortions caused by hidden data embedding. Transformer encoders modeled long-range temporal dependencies in the audio representations.

Model	Accuracy	Precision	Recall	F1-Score
CNN Only	88.9%	88.2%	89.5%	88.8%
Transformer Only	87.4%	86.9%	88.1%	87.5%
Hybrid CNN-Transformer	92.7%	92.1%	93.3%	92.7%

Table 2. Audio Steganalysis Results

Figure 3. Confusion matrix for audio steganalysis.

The audio results shown in Figure 3 and Table 2 summarize that the proposed framework effectively identified hidden information across various embedding methods. The hybrid architecture achieved higher recall values, showing stronger sensitivity to stego-audio detection.

The comparatively lower performance relative to image steganalysis is expected. Audio steganography usually leads to smaller perceptual changes and temporal variability, which makes detection harder.

4.3 Video Steganalysis Results:

After generating stego-video samples (in the spatial and frequency domain embedding methods), the video steganalysis module was tested on HMDB51, UCF-101, and Kinetics-400 datasets.

Frame-level spatial artifacts were extracted from the ResNet-18 encoder and learned temporal relations of consecutive frames by Transformer encoders.

Model	Accuracy	Precision	Recall	F1-Score
CNN Only	86.7%	85.9%	87.3%	86.6%
Transformer Only	85.8%	85.15%	86.4%	85.7%
Hybrid CNN-Transformer	90.8%	90.1%	91.4%	90.7%

Table 3. Video Steganalysis Results

Figure 4. Confusion matrix for video steganalysis.

Among all modalities, video steganalysis achieved comparatively lower accuracy (as compared to image and audio modalities) due to its additional temporal complexity, compression artifacts, motion variations, and redundancy with frames in a sequence.

Despite these challenges, the hybrid model maintained stable classification performance and successfully captured both spatial and temporal embedding patterns.

4.4 Multimodal Fusion Results:

The proposed late fusion strategy combined feature embeddings extracted from image, audio, and video modalities to generate a unified multimedia steganalysis prediction.

Model	Accuracy	Precision	Recall	F1-Score
Images only	95.2%	94.8%	95.7%	95.2%
Audio only	92.7%	92.1%	93.3%	92.7%
Video only	90.8%	90.1%	91.4%	90.7%
Multimodal Fusion	96.4%	96.0%	96.8%	96.4%

Table 4. Multimodal Fusion Results

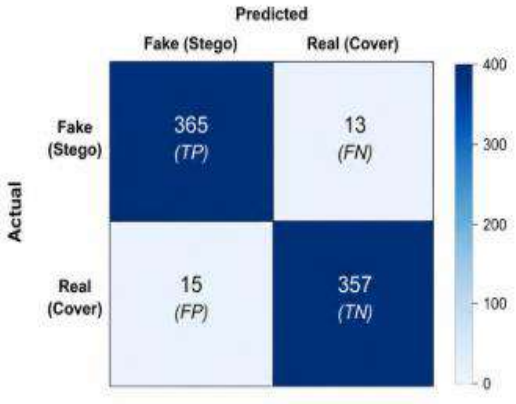

Figure 5. Confusion matrix multimodal fusion steganalysis.

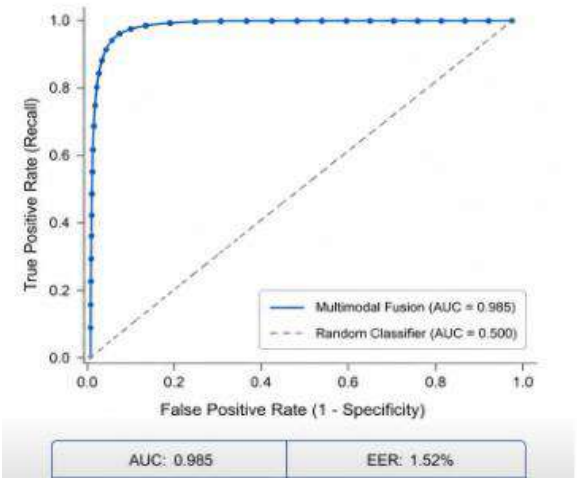

Figure 6. ROC curve for multimodal fusion steganalysis.

The multimodal late-fusion framework achieved the highest overall performance, as shown in Figure 5 and Table 4. The fusion process improved robustness because different modalities contributed complementary embedding characteristics.

The results demonstrate that combining multimodal representations can reduce modality-specific weaknesses and improve overall detection reliability.

V. Conclusion

This paper presented a unified multimedia steganalysis framework based on a Hybrid CNN–Transformer architecture for detecting hidden information in image, audio, and video data. The proposed framework combined the local feature extraction capability of Convolutional Neural Networks (CNNs) with the global contextual learning capability of Transformer encoders to effectively identify spatial, spectral, and temporal steganographic artifacts across multiple multimedia modalities.

The framework was evaluated using publicly available datasets including BOWSBASE, BOWS2, TIMIT, ESC-50, LibriSpeech, HMDB51, UCF-101, and Kinetics-400. Experimental results demonstrated that the proposed Hybrid CNN–Transformer model achieved improved detection performance compared to standalone CNN and Transformer architectures, providing better feature representation, multimodal generalization, and classification robustness.

Despite achieving strong performance, the proposed framework has certain limitations, including high computational complexity, increased training time for video processing, and dependency on high-performance GPU resources. Furthermore, the current evaluation is primarily based on publicly available datasets, which may not completely represent real-world adversarial steganography scenarios.

Overall, the proposed framework demonstrates the effectiveness of combining CNN and Transformer architectures for multimedia steganalysis and provides a promising direction for future research in intelligent cybersecurity and multimedia forensics applications.

Future Work:

Future research should address the efficiency, scalability, and robustness of the proposed multimedia steganalysis framework. A promising direction involves developing lightweight Hybrid CNN–Transformer architectures to reduce computational complexity and facilitate real-time deployment on resource-constrained systems. Additionally, optimization techniques such as model pruning, quantization, and knowledge distillation could further enhance inference efficiency while maintaining detection accuracy.

References

[1] De La Croix, N. J., Ahmad, T., & Han, F. (2024). Comprehensive survey on image steganalysis using deep learning. *array*, *22*, 100353. https://doi.org/10.1016/j.array.2024.100353

[2] Anderson, R. J., & Petitcolas, F. A. (2002). On the limits of steganography. *IEEE Journal on selected areas in communications*, *16*(4), 474-481. https://doi.org/10.1109/49.668971

[3] Fridrich, J. (2009). *Steganography in digital media: principles, algorithms, and applications*. Cambridge university press. https://doi.org/10.1017/cbo9781139192903.006

[4] Shankar, D. D., & Azhakath, A. S. (2021). Minor blind feature based Steganalysis for calibrated JPEG images with cross validation and classification using SVM and SVM-PSO. *Multimedia Tools and Applications*, *80*(3), 4073-4092. https://doi.org/10.1007/s11042-020-09820-7

[5 Kodovsky, J., Fridrich, J., & Holub, V. (2011). Ensemble classifiers for steganalysis of digital media. *IEEE Transactions on information forensics and security*, *7*(2), 432-444. https://doi.org/10.1109/tifs.2011.2175919

[6] Pevný, T., Bas, P., & Fridrich, J. (2009, September). Steganalysis by subtractive pixel adjacency matrix. In *Proceedings of the 11th ACM workshop on Multimedia and security* (pp. 75-84). https://doi.org/10.1145/1597817.1597831

[7] Qian, Y., Dong, J., Wang, W., & Tan, T. (2015, March). Deep learning for steganalysis via convolutional neural networks. In *Media Watermarking, Security, and Forensics 2015* (Vol. 9409, pp. 171-180). SPIE. https://doi.org/10.1117/12.2083479

[8] Yedroudj, M., Comby, F., & Chaumont, M. (2018, April). Yedroudj-net: An efficient CNN for spatial steganalysis. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2092-2096). IEEE. https://doi.org/10.1109/icassp.2018.8461438

[9] Xu, G., Wu, H. Z., & Shi, Y. Q. (2016). Structural design of convolutional neural networks for steganalysis. *IEEE Signal Processing Letters*, *23*(5), 708-712. https://doi.org/10.1109/lsp.2016.2548421

[10] Boroumand, M., Chen, M., & Fridrich, J. (2018). Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, *14*(5), 1181-1193. https://doi.org/10.1109/tifs.2018.2871749

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*. https://doi.org/10.48550/arXiv.1706.03762

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022). https://doi.org/10.1109/iccv48922.2021.00986

[14] Luo, G., Wei, P., Zhu, S., Zhang, X., Qian, Z., & Li, S. (2022, May). Image steganalysis with convolutional vision transformer. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3089-3093). IEEE. https://doi.org/10.1109/icassp43922.2022.9747091

[15] Bravo-Ortiz, M. A., Mercado-Ruiz, E., Villa-Pulgarin, J. P., Hormaza-Cardona, C. A., Quiñones-Arredondo, S., Arteaga-Arteaga, H. B., ... & Tabares-Soto, R. (2024). CVTStego-Net: A convolutional vision transformer architecture for spatial image steganalysis. *Journal of Information Security and Applications*, *81*, 103695. https://doi.org/10.1016/j.jisa.2023.103695

[16] Sedighi, V., Cogranne, R., & Fridrich, J. (2015). Content-adaptive steganography by minimizing statistical detectability. *IEEE transactions on information forensics and security*, *11*(2), 221-234. https://doi.org/10.1109/tifs.2015.2486744

[17] Fridrich, J., & Kodovsky, J. (2012). Rich models for steganalysis of digital images. *IEEE Transactions on information Forensics and Security*, *7*(3), 868-882. https://doi.org/10.1109/tifs.2012.2190402

[18] Allen, J. D., Liu, X., Mayron, L. M., & Mio, W. (2010, April). On generalization of performance of classifiers for steganalysis. In *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research* (pp. 1-1). https://doi.org/10.1145/1852666.1852744

[19] Zhang, Z., Yi, X., & Zhao, X. (2019, November). Improving audio steganalysis using deep residual networks. In *International Workshop on Digital Watermarking* (pp. 57-70). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-43575-2_5

[20] Li, Z., Jiang, X., Dong, Y., Meng, L., & Sun, T. (2022). An anti-steganalysis HEVC video steganography with high performance based on CNN and PU partition modes. *IEEE Transactions on Dependable and Secure Computing*, *20*(1), 606-619. https://doi.org/10.1109/tdsc.2022.3140899

[21] Reinel, T. S., Brayan, A. A. H., Alejandro, B. O. M., Alejandro, M. R., Daniel, A. G., Alejandro, A. G. J., ... & Raul, R. P. (2021). GBRAS-Net: a convolutional neural network architecture for spatial image steganalysis. *IEEE Access*, *9*, 14340-14350. https://doi.org/10.1109/access.2021.3052494

[22] Zhang, C., Jiang, S., & Chen, Z. (2024). TENet: leveraging transformer encoders for steganalysis of QIM steganography in VoIP speech streams. *Multimedia tools and applications*, *83*(19), 57107-57138. https://doi.org/10.1007/s11042-023-17802-8

[23] Wei, K. K., Luo, W. Q., Tan, S. Q., & Huang, J. W. (2025). Ctnet: A convolutional transformer network for color image steganalysis. *Journal of Computer Science and Technology*, *40*(2), 413-427. https://doi.org/10.1007/s11390-023-3006-3

[24] WANG, J., Yan, H., & Gu, J. (2025). Image steganalysis based on CNN-Transformer. *Journal of Shenzhen University (Science and Engineering)*, *42*(2), 233-241. https://doi.org/10.3724/sp.j.1249.2025.02233

[25] Peng, J., Liao, Y., & Tang, S. (2023, November). VoIP Steganalysis Using Shallow Multiscale Convolution and Transformer. In *International Conference on Security and Privacy in New Computing Environments* (pp. 326-351). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-73699-5_23