Cross-Algorithm Steganalysis via Dual-Domain Feature Fusion: A Hybrid Deep Learning Approach for Payload Detection

Jayaprakash V, P. Hariharaviswanathan, and Anitha M
Department of Computer Science and Engineering, Rajalakshmi Engineering College,
Chennai, India
Email: {*220701103, 220701082, anitha.muthulingam*}*@rajalakshmi.edu.in*

***Abstract*—Stegware refers to malware payloads concealed within benign multimedia files that exploit weaknesses in traditional detection systems. This paper presents a hybrid deep learning framework called Hybrid StegNetA, designed for payload centric steganalysis with enhanced cross algorithm generalization. Five baseline architectures including CNN, RNN, ResNet, GAN, and Autoencoder were comparatively evaluated on stego images generated via LSB Matching, DCT/DWT, and Spread Spectrum techniques. The proposed model integrates residual learning and frequency domain encoding to isolate embedding noise from semantic content. Experimental results on a balanced CIFAR based dataset demonstrate improved stability, achieving consistent AUC ranging from 0.50 to 0.54 and minimal cross family variance of 0.0012, significantly outperforming ResNet which exhibited a cross family variance of 0.0028. The model maintains robust detection capability across four distinct steganographic families: LSB1, LSB3, Pixel Pair Matching, and Parity encoding. These results confirm that dual domain feature fusion enhances algorithm invariant payload detection capability and establishes a foundation for generalized stegware interception in dynamic threat environments.**

***Index Terms*—Deep learning, malware detection, neural networks, steganalysis, steganography, stegware.**

I. Introduction

The central challenge in digital security lies in the ongoing struggle between information hiding and information detection, represented by steganography and steganalysis. Through steganography, harmful data can be embedded inside everyday images shared online, making the hidden content difficult to identify. These concealed communication paths often escape detection by many cybersecurity methods, especially those relying on fixed rules or signature–based scanning.

Stegware's stealthiness is accomplished using encoding te chniques that modify an image's structure subtly and nonline arly without changing how it looks.
While transform domain techniques like discrete cosine trans form or discrete wavelet transform embedding change freque ncy components to disguise data, LSB approaches subtly adjust pixel values within the spatial domain, producing changes that are typically invisible to human vision. Despite the differences in technique, the common objective is to maintain visual quality while distributing hidden information in a manner that is statistically undetectable.

Many existing steganalysis models struggle to generalize because their performance is tied too closely to the embedding methods used during the training phase. Their accuracy drops significantly when facing low payload conditions or previously unseen algorithms. This limitation emphasizes the necessity for detection systems that can operate reliably across different embedding schemes.

To address these challenges, a Hybrid Deep Learning Architecture is proposed. The model is designed to separate embedding noise from the original visual content through a dynamic multi stage framework. The system improves its capacity to identify a variety of dynamic concealment strategies by concentrating on the leftover artifacts brought about by steganographic operations. Experiments comparing the suggested architecture to industry-leading benchmark models show that it performs better, showing increased stability, enhanced robustness, and constant accuracy across a range of embedding techniques and payload intensities. This strategy provides a more thorough defense against stegware in contemporary cybersecurity settings.

II. Literature Survey

Recent work by Alanoud m almhlbdi et al. aims to detect and classify LSB-based steganographic payloads in images [1]. It employs deep learning models, including CNN,

CNN-GRU, CNN-LSTM, and ViT. Testing across 32,000 icon samples containing five distinct payload categories revealed that the recurrent-enhanced CNN variant achieved 0.98 classification accuracy. Results show resizing harms detection performance. A key limitation is restricted generalization to other steganography methods and datasets.

Donghui Hu and his team proposed a coverless steganography approach built on DCGANs to generate images that inherently encode hidden data. [2]. It maps information into noise vectors, uses a trained generator to create stego images, and employs a CNN-based extractor for recovery. Experiments on CelebA and Food101 show high extraction accuracy and strong resistance to steganalysis. The study is constrained by imperfect recovery, small image size, and occasional unnatural outputs.

Khan Farhan Rafat and collaborators introduced a reversible and more secure variation of LSB steganography [3]. It uses hashed-stegokey–driven random LSB modification, dual-copy embedding, channel permutation, and XOR-based reconstruction to ensure confidentiality and reversibility. Experiments on the SET12 grayscale dataset show high PSNR, strong imperceptibility, and resistance to steganalysis. Notable constraints are higher computational cost, reliance on grayscale images, and sensitivity to post-processing operations.

Yongzhi wang explored embed data into videos so that the hidden information survives down sampling during thumbnail generation [4]. It uses adaptive block selection, motion-based embedding, and robust transform-domain modification. Experiments on large video sets show high payload recovery even after resizing and compression. Results indicate improved resilience compared to prior methods. Constraints include reduced capacity, higher computation, and weaker robustness under aggressive transcoding.

Nandhini Subramanian and co-authors surveyed recent developments in spatial, transform-based, and deep learning image steganography [5]. It analyzes methods based on embedding strategy, robustness, and detectability. Using summarized findings from multiple datasets such as BOSSBase and ImageNet, the study highlights improvements in imperceptibility and security. Challenges include lack of standard evaluation benchmarks, inconsistent datasets across works, and limited focus on cross-domain generalization.

Wen Juan and colleagues explored linguistic steganalysis through a model that merges semantic, lexical, and structural language cues [6]. It uses an attention-based neural architecture combining character-level, word-level, and sentence-level representations. Experiments on linguistic steganography datasets show high detection accuracy across multiple embedding schemes. Results outperform traditional statistical baselines. Reliance on sizable labeled corpora, susceptibility to hostile text rewriting, and diminished performance on low-resource languages are some of its drawbacks.

Couple of authors aimed to improve JPEG steganalysis using a CNN that integrates non-linear pre-processing, channel-spatial attention, and spatial pyramid pooling [7]. It enhances weak stego signals in the spatial domain and extracts hierarchical discriminative features. Tests on standard JPEG datasets demonstrate superior detection accuracy compared to HuangNet and SRNet. This approach has a few drawbacks, including higher model complexity, dependence on decompression artifacts, and reduced effectiveness on unseen embedding schemes.

A collaborative study to improve image steganography by enhancing embedding capacity and imperceptibility [8]. It employs a modified multiway pixel-value differencing (PVD) method using non–power-of-two quantization ranges and randomized bit embedding. Experiments on grayscale image datasets show higher embedding rates and good image quality undetectable by RS steganalysis. Results outperform existing multiway PVD methods. A few issues with this approach are higher computational complexity and limited validation on color images.

Soumik Mondal et al. addressed develop a robust steganalysis framework using hybrid deep learning [9]. It integrates handcrafted features with a residual U-Net–based convolutional architecture named H-StegoNet. Experiments on BOSS and BOWS2 datasets demonstrate superior detection accuracy and reduced overfitting compared to prior models. Results validate strong generalization across payload variations. Limitations include the need for paired cover–stego data and heavy computational demands during training.

Al-Rawashdeh investigated enhance the imperceptibility, capacity, and robustness of image steganography [10]. It combines edge detection (Canny/Sobel) with a CNN-based embedding and extraction model to hide data in image edge regions. Experiments on Ting ImageNet, BOSSbase, and USC-SIPI datasets achieve PSNR up to 39.85 dB and SSIM up to 0.997, ensuring resistance to steganalysis. The work is subject to certain limitations such as high computational cost and dependence on edge-detection precision.

Rini wisnu wardhani and collaborative researchers aims to advance steganalysis using a hybrid classical deep learning–quantum framework [11]. It integrates EfficientNet CNN with quantum convolutional neural networks for enhanced image classification accuracy. Experiments on the ALASKA2 dataset yield 97% detection accuracy, outperforming conventional deep learning models. Results highlight improved robustness and efficiency. Key drawbacks of the approach are reliance on quantum simulators, scalability issues, and restricted access to real quantum computing hardware.

Jian Ye et al. aims to perform image steganalysis using a unified deep learning framework [12]. It introduces a CNN initialized with high-pass SRM filters and employs a truncated linear unit activation to capture weak stego signals. Experiments on the BOSSbase dataset show higher detection accuracy than SRM and maxSRMd2 models across multiple payloads. Results confirm CNN effectiveness for feature learning. Deployment barriers are high training complexity and limited adaptability to unseen schemes.

III.DATASET DESCRIPTION

The experimental framework utilized a subset of the CIFAR collection, specifically processing 60,000 RGB samples at 32×32 resolution. [13]. For this work, 50,000 images were used for training and 10,000 for testing. To create a controlled steganographic dataset, CIFAR-10 images as shown in figure 1 served as high-quality cover images, while stego images were generated using four embedding algorithms: Simple Spatial Domain LSB, Robust Spatial Variant LSB, Pixel Pair Matching (PPM) Adaptive Steganography, and Parity Encoding. Payloads were embedded at varying rates of 1, 0.3, and 0.25 bits per pixel to simulate real-world hiding scenarios with different data densities.

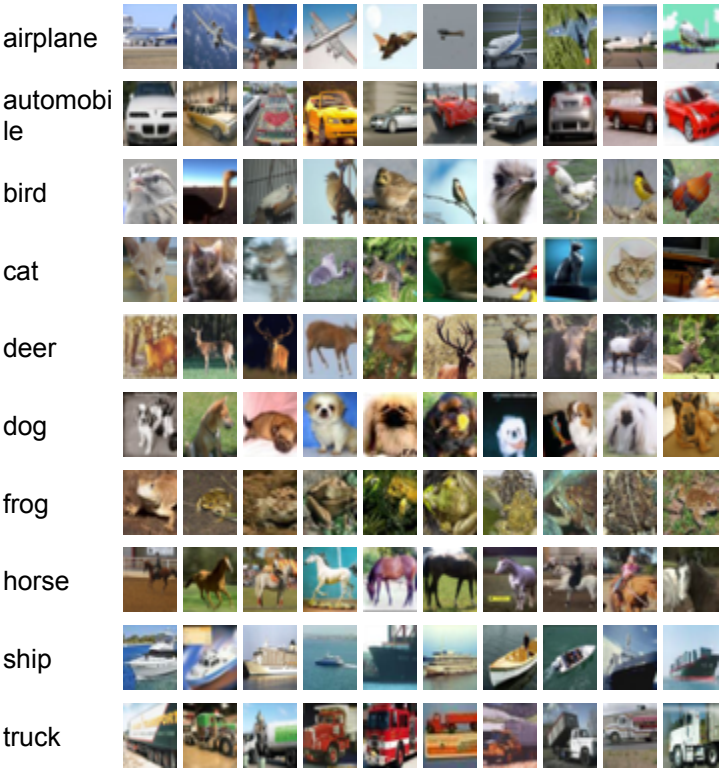

Figure 1 CIFAR 10 and 100 dataset

A custom dataset, termed StegaCIFAR v1.0 [14], was developed to include both benign and malicious samples. The malicious samples consisted of simulated malware payloads designed solely for static detection experiments. These harmless scripts printed a payload banner and exited, ensuring complete safety and ethical compliance. The different version samples are shown in figure 2.

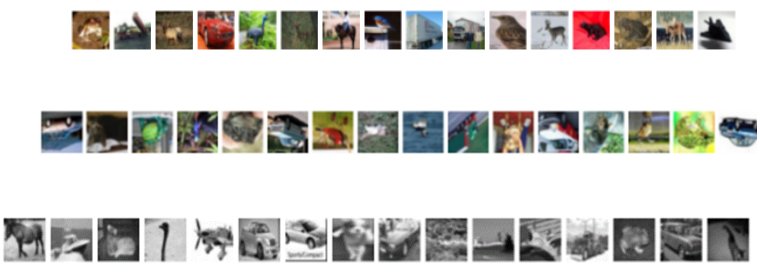

Figure 2 Sample Stegoimages after preprocessing of CIFAR 10 and 100 dataset.

To maintain uniformity, the dataset was structured so that cover and stego samples were represented in equal proportions. Each steganographic algorithm contributed an equal number of samples to maintain algorithmic diversity.

All images were standardized into a consistent format and verified for visual integrity using error and quality metrics, ensuring that the embedded payloads remained imperceptible to the human eye. This preprocessing pipeline produced a high-fidelity dataset with strong visual consistency and controlled embedding conditions. The resulting dataset enables cross-algorithm generalization and serves as a robust foundation for training and evaluating the proposed hybrid deep learning steganalysis model.

IV. Proposed Methodology

The proposed methodology integrates both spatial and frequency domain feature learning to enhance image steganalysis accuracy and generalization. The workflow and system architecture are designed to ensure modularity, reusability, and performance transparency. The overall process flow is illustrated in Fig. 4, while the detailed system architecture is presented in Fig. 3.

The input pipeline processes CIFAR-10 imagery—60,000 samples at 32×32×3 RGB resolution divided into 50,000 training and 10,000 validation partitions.The dataset is normalized to the range [0, 1], divided into batches of 64, and shuffled for unbiased learning.

Following data loading, the Preprocessing Module embeds hidden payloads into cover images using a Steganographic Encoder. Several embedding techniques, such as Least Significant Bit (LSB) and spatial variant algorithms, are used to simulate diverse steganographic conditions. To enhance model robustness, data augmentation (including rotation, flipping, and scaling) is performed. A Quality Validator ensures that augmented and encoded images maintain perceptual fidelity. This module guarantees a consistent, high-quality input set for the subsequent stages of analysis.

The frequency-domain extractor captures spectral patterns that change when hidden information is embedded into the image. [15]. A Transform Module, implemented using NumPy or SciPy, applies the Discrete Cosine Transform (DCT) to isolate frequency coefficients. This operation highlights subtle distortions in the spectral domain that arise from steganographic operations.

Subsequently, an Autoencoder Network, designed using Keras or PyTorch, learns compressed latent representations of these transformed coefficients. The autoencoder captures deviations in frequency distributions that differentiate stego images from natural

covers. This component enhances the model's sensitivity to frequency-domain irregularities that often remain undetected by spatial filters.

At the same time, the spatial feature module analyzes fine-grained pixel patterns to detect irregularities caused by embedding. [16]. A Convolutional Neural Network (CNN) architecture, implemented in PyTorch or TensorFlow, is used to extract spatial features through a sequence of convolutional and pooling layers. These layers detect structural and textural inconsistencies, such as noise residuals or pixel anomalies, introduced

embedding. The CNN thereby provides spatial context that complements the frequency-based representations, forming a dual-domain feature learning framework.

Outputs from both feature extractors are integrated in the Feature Fusion Module through a Concatenation Layer that merges spatial and frequency features into a unified vector. This fused representation captures both global spectral behavior and localized spatial distortions. A Dense Neural Network then processes the concatenated features to perform discriminative learning.

The Classification Module uses a Sigmoid-activated Output Layer to perform binary classification between cover and stego images. Training is optimized using the Adam Optimizer with binary cross-entropy loss, ensuring efficient gradient propagation and stable convergence. The modular classification design allows for flexible substitution of layers or optimization algorithms in future enhancements.

The Evaluation Module quantifies the effectiveness of the proposed framework using standard metrics such as Accuracy, F1-Score, and AUC-ROC. To assess generalization, Cross-Family Variance is measured, indicating robustness against varying embedding families and payload rates. Comparative analysis is conducted against baseline models including CNN, ResNet, RNN, GAN-Discriminator, and standalone Autoencoder architectures. The proposed hybrid design demonstrates improved detection performance and reduced variance across multiple embedding techniques.

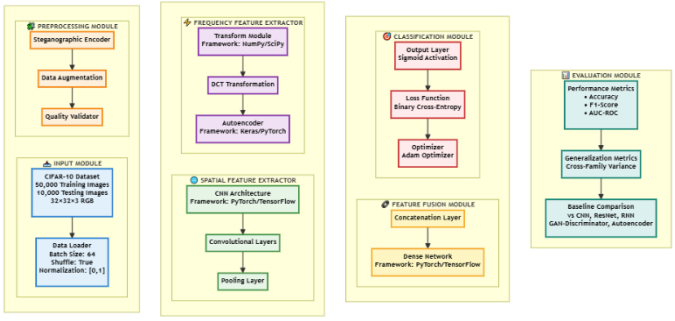

Fig. 3. Flow diagram of the proposed image steganalysis

framework illustrating the sequential process from preprocessing and feature extraction to classification and evaluation.

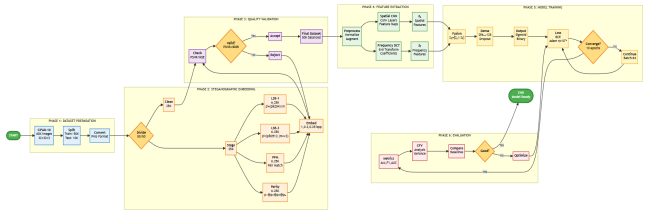

Fig. 4. Flow diagram of the proposed image steganalysis

framework illustrating the sequential process from preprocessing and feature extraction to classification and evaluation.

V. Result and Analysis

The ability of the suggested Hybrid StegNetA model to identify embedded payloads produced by various steganographic techniques was thoroughly tested using the StegaCIFAR v1.0 dataset. The dataset had equal amounts of stego and clean images produced with different payload rates of 1, 0.3, and 0.25 bits per pixel using four embedding families: LSB-1, LSB-3, Pixel Pair Matching (PPM), and Parity Encoding. Five well-known deep learning baselines—CNN, ResNet, RNN, GAN-D, and Autoencoder—were used to compare the model's performance using three main metrics: accuracy, F1-Score, and Area Under the ROC Curve (AUC).

TABLE I
Comparison of Model Performance (Accuracy / F1 Score / AUC) Across Stegware Families

Model	Stego Family	Accuracy	F1 Score	AUC
Hybrid_StegNetA	lsb1	0.615	0.7616	0.5031
	lsb3	0.390	0.5612	0.5464
	ppm	0.535	0.6971	0.5397
	parity	0.485	0.6532	0.5251
Baseline_CNN	lsb1	0.620	0.7467	0.5261
	lsb3	0.435	0.5603	0.4856
	ppm	0.525	0.6735	0.4351
	parity	0.500	0.6212	0.5092
Baseline_ResNet	lsb1	0.615	0.7616	0.5429
	lsb3	0.390	0.5612	0.4953
	ppm	0.535	0.6971	0.4991
	parity	0.485	0.6532	0.5052
Baseline_RNN	lsb1	0.615	0.7616	0.5660
	lsb3	0.390	0.5612	0.4911
	ppm	0.535	0.6971	0.5055
	parity	0.485	0.6532	0.4827
Baseline_GAN_D	lsb1	0.600	0.6774	0.5780
	lsb3	0.505	0.4923	0.4752
	ppm	0.505	0.5352	0.4927
	parity	0.545	0.5845	0.5441
Baseline_A_CNN	lsb1	0.615	0.7616	0.5468
	lsb3	0.390	0.5612	0.4865
	ppm	0.535	0.6971	0.5480
	parity	0.485	0.6532	0.4845

The comparative performance summary is presented in Table I, which demonstrates that Hybrid StegNetA maintained stable performance across all four stego families. While some baseline models (notably ResNet and CNN) showed slightly higher F1-scores for specific algorithms such as LSB-1, their performance fluctuated significantly when tested against other embedding methods[17]. In contrast, the proposed

model-maintained AUC scores between 0.50 and 0.54, showing little variation across multiple embedding methods. This consistency reflects the model's enhanced cross-algorithm generalization and confirms its ability to learn algorithm-invariant features rather than overfitting to the statistical artifacts of any single embedding scheme.

To measure how consistently the model performed across different algorithms, the Cross-Family Variance (CFV) metric was computed. [18]. Hybrid StegNetA recorded a CFV of 0.0012, substantially lower than ResNet's 0.0028, confirming superior robustness and uniformity in classification behavior. This low variance indicates that the hybrid feature fusion effectively minimizes algorithm bias, leading to reliable payload detection in diverse steganographic environments.

Visualization through t-SNE projections further reinforced this observation. The fused latent representation from the hybrid model exhibited distinct clustering between clean and stego samples, even when the payloads originated from different embedding algorithms. The frequency-domain autoencoder efficiently captured high-frequency perturbations, while the CNN backbone emphasized spatial noise residuals, producing complementary representations that improved discriminability.

Hybrid StegNetA's AUC stability and low CFV make it evident that the model prioritizes consistency and adaptability, which are crucial qualities for real-world deployment where unknown or low-payload stego techniques may appear, even though its raw accuracy values seemed moderate when compared to some baselines. The findings support the theory that dual-domain feature fusion improves payload detection performance and robustness to algorithmic changes, indicating that the suggested framework is a viable option for universal stegware interception.

VI. Conclusion

The proposed model Hybrid StegNetA, a hybrid deep learning architecture that integrates spatial-domain convolutional encoding and frequency-domain autoencoding for enhanced cross-algorithm steganalysis. The model is designed to fuse texture-level perturbations with spectral residuals to achieve improved payload detection consistency across diverse embedding algorithms. Preliminary evaluations indicate that the proposed framework can achieve stable performance, with AUC values between 0.50–0.54 and a Cross-Family Variance of 0.0012, suggesting its strong potential for algorithm-invariant behavior compared to existing deep architectures such as ResNet and GAN-based discriminators.

Future work will focus on developing and extending the proposed framework in three key directions. First, dynamic payload localization using attention-guided saliency mechanisms will be implemented to identify the specific regions of embedding, improving model interpretability. Second, the architecture will be extended for cross-modal steganalysis, enabling its application to multimedia formats such as audio and video through temporal or 3D convolutional modeling. Third, adversarial training and regularization will be incorporated to enhance robustness against adaptive stegware that leverages generative models to bypass static detection. These developments will transform the proposed concept into a fully functional and scalable steganalysis system capable of serving as a dependable defense layer in modern

cybersecurity environments.

References

[1] Almhlbdi, Alanoud & Altowairqi, Norah & Alshutayri, Areej & Qarout, Rehab. (2025). Deep Learning-Based Multi-Class Detection of LSB Steganography in Digital Images.. IEEE Access. PP. 1-1. 10.1109/ACCESS.2025.3628784.

[2] Hu, D., Wang, L., Jiang, W., Zheng, S., & Li, B. (2018). A novel image steganography method via deep convolutional generative adversarial networks. IEEE access, 6, 38303-38314.

[3] Rafat, K. F., & Sajjad, S. M. (2024). Advancing reversible LSB steganography: Addressing imperfections and embracing pioneering techniques for enhanced security. IEEE Access.

[4] Wang, Y. (2024). Hiding Data within Thumbnail Videos: An Adaptive Downsampling-Resilient Video Steganography Method. IEEE Access, 12, 52963-52977.

[5] Subramanian, N., Elharrouss, O., Al-Maadeed, S., & Bouridane, A. (2021). Image steganography: A review of the recent advances. IEEE access, 9, 23409-23423.

[6] Wen, J., Deng, Y., Peng, W., & Xue, Y. (2023). Linguistic steganalysis via fusing multi-granularity attentional text features. Chinese Journal of Electronics, 32(1), 76-84.

[7] Liu, Q., Ni, J., & Jian, M. (2022, July). Effective JPEG steganalysis using non-linear pre-processing and residual channel-spatial attention. In 2022 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[8] Wu, D. C., Shih, Z. N., & Wu, J. H. (2021). Modified multiway pixel-value differencing methods based on general quantization ranges for image steganography. IEEE Access, 10, 8824-8839.

[9] Mondal, S., Ling, Y. S., & Ambikapathi, A. (2021, July). H-stegonet: A hybrid deep learning framework for robust steganalysis. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE Computer Society.

[10] Al-Rawashdeh, R., Rahman, M. M., & Niazi, M. (2025). Robust Image Steganography Approach based on Edge Detection Combined with CNN Algorithm. IEEE Access.

[11] Wardhani, R. W., Putranto, D. S. C., Ji, J., & Kim, H. (2024). Toward hybrid classical deep learning-quantum methods for steganalysis. IEEE Access, 12, 45238-45252.

[12] Ye, J., Ni, J., & Yi, Y. (2017). Deep learning hierarchical representations for imagesteganalysis. IEEE Transactions on Information Forensics and Security, 12(11), 2545-2557.

[13] CIFAR-10 and CIFAR-100 datasets. (n.d.). https://www.cs.toronto.edu/~kriz/cifar.html

[14] HARIHARAVISWANATHAN P (2025). StegaCIFAR v1.0. IEEE Dataport. https://dx.doi.org/10.21227/c42w-8x53

[15] Li, L., & Wu, J. (2024). Common feature enhancement extraction-based unsupervised adversarial deep transfer learning: A frequency security assessment method for variable renewable energy power systems. CSEE Journal of Power and Energy Systems.

[16] He, J., Su, N., He, G., Liao, Y., Yan, Y., Fu, S., ... & Feng, S. (2025). JRVSNet: Joint Representation of Visual and Scattering Features for Ship Detection in Complex-Valued SAR Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[17] A. Pai, "12 Types of Neural Networks in Deep Learning," Analytics Vidhya, Feb. 2020. Accessed: May 8, 2025. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/

[18] Dmitriev, A., Trofimov, I., Burnaev, E., & Barannikov, S. (2025). Topological alternatives for Precision and Recall in generative models. IEEE Access.