Unveiling the Power of Deep Learning in Steganography Classifications

Saravana Gokul G
Department of CSE
Rajalakshmi Engineering College
Chennai, India
sharavana.ssg@gmail.com

Deepak Kumar K
Department of CSE
Rajalakshmi Engineering College
Chennai, India
kdeepak.srmit@gmail.com

Senthil Pandi S
Department of CSE
Rajalakshmi Engineering College
Chennai, India
mailtosenthil.ks@gmail.com

Kumar P
Department of CSE
Rajalakshmi Engineering College
Chennai, India
kumar@rajalakshmi.edu.in

Neha M U
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701178@rajalakshmi.edu.in

Monika S
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701166@rajalakshmi.edu.in

***Abstract*—In today's digital landscape, the implementation of steganography to hide sensitive information within images has become increasingly sophisticated, posing new challenges for secure data communication. This paper presents a deep learning technique specifically focused on the detection of stego-images—images with hidden messages—utilizing the VGG16 Convolutional Neural Network (CNN) architecture. The detection model is designed to identify subtle pixel-level modifications characteristic of steganographic methods, particularly the Pixel Value Differencing (PVD) and the Least Significant Bit (LSB) techniques, which are widely used for data embedding. The VGG16 model was trained on a carefully curated dataset of stego and non-stego images, our approach effectively learns to distinguish between the two with high accuracy. Through extensive evaluation, we demonstrate that our model achieves robust performance in terms of recall, precision, accuracy and F1 score, underscoring its reliability in classifying stego-images. This work contributes a powerful tool for enhancing secure digital communication, offering a systematic method to detect hidden information within images, thereby addressing a critical need for modern data security practices.**

Keywords—Steganography Detection, Deep Learning, Convolutional Neural Networks (CNN), Least Significant Bit (LSB), VGG16 Architecture, Image Classification, Pixel Value Differencing (PVD).

I. INTRODUCTION

With the rapid growth of digital communication, the need for secure data transmission has become increasingly critical. Steganography, the technique of hiding information within digital media, offers a subtle method for embedding messages without drawing attention to their presence, in contrast to traditional encryption methods. By embedding information within images, steganography conceals sensitive data within pixel variations, often undetectable to the human eye. However, as steganographic methods like Pixel Value Differencing (PVD) and Least Significant Bit (LSB) become more sophisticated, so too do the tools required to detect them, presenting challenges in maintaining secure and covert communication channels. In response to these challenges, deep learning techniques have shown a great promise in correctly identifying stego-images—images that contain hidden information. This paper concentrates on the application of Convolutional Neural Networks (CNNs), specifically the VGG16 architecture, for the detection of steganographic content within images. Leveraging VGG16's advanced feature extraction capabilities, the model is trained to distinguish stego-images from non-stego-images by identifying the subtle pixel-level alterations introduced by LSB and PVD methods. This paper presents a comprehensive deep learning framework designed to improve the accuracy of steganography detection. We assess the model's performance, ultimately contributing a reliable tool for secure digital communication. The following sections detail the dataset, model architecture, experimental results, and evaluation metrics, offering insights into the impact of deep learning on advancing steganography detection.

II. LITERATURE SURVEY

Atique ur Rehman et al. [1] This study introduces a deep learning encoder-decoder model that embeds images into cover images using a unique loss function, achieving high data capacity and image quality but requiring significant computational power. Jinyuan Tao et al. [2] This review examines recent advancements in deep learning for image steganography, categorizing techniques as traditional, CNN-based, and GAN-based, and focusing on theoretical insights rather than practical applications. Kumar P et al. [3] proposed a method combining CNN and edge-based segmentation was proposed to enhance medicinal plant identification, achieving higher accuracy than traditional approaches. Kumar P et al. [4] A face mask detection model utilizing data augmentation techniques was proposed to improve generalization and accuracy by addressing input variability. Donghui Hu et al. [5] This research employs Deep Convolutional GANs to generate secure carrier images for hidden data, but it faces high computational demands and limitations in data concealment capacity. Kevin A. Zhang et al. [6] Stegano GAN embeds binary data in images via GANs, achieving 4.4 bits per pixel while evading detection, but requires complex training processes and significant computational resources. Weixuan Tang et al. [7] An adversarial embedding technique modifies image elements to create stego images that evade CNN-based detection, though it may introduce detectable artifacts. Nandhini Subramanian et al. [8] This paper discusses using CNNs to enhance detection of hidden data in steganography through global statistical constraints and transfer learning, necessitating large datasets and computational power. Jiaohua Qin et al. [9] GANs are used to directly generate stego images from secret information, enhancing security but requiring substantial computational resources and training efforts.Ying Zou et al. [10] The technique for embedding data in compressed images focuses on JPEG robustness via coefficient adjustment but may be less effective with other processing methods. Jiaohua Qin et al. [11] This survey explores coverless steganography, concealing information without altering cover images. It highlights lower data capacity compared to traditional methods. Jiwen Yu et al. [12] Diffusion models improve image steganography's security and robustness, preserving high visual quality but requiring advanced knowledge and significant computational resources.Alejandro Martín et al. [13] This method employs GANs for LSB-based embedding of secret messages, though high computational demands may limit its effectiveness across scenarios. Mikołaj Płachta et al. [14] Various machine learning algorithms are investigated for detecting JPEG steganographic modifications, achieving high detection rates, but accuracy varies by algorithm. Supriadi Rustad et al. [15] An adaptive steganography method using inverted LSB enhances imperceptibility but is computationally intensive and dependent on cover image characteristics. Sabyasachi Pramanik et al. [16] A modified Genetic Algorithm improves image steganography's security and efficiency, enhancing embedding quality but requiring considerable computational resources. Vijay Kumar et al. [17] This review focuses on recent developments in deep learning for image steganography, particularly GANs, discussing strengths and weaknesses while suggesting future research directions. Pratik D. Shah et al. [18] A genetic algorithm optimizes secret data embedding into LSBs of cover images, enhancing security and quality but being computationally intensive. M.K. Shyla et al. [19] A GA-based method selects suitable cover images for secret data embedding, improving performance but also being resource-intensive. Jiaohua Qin et al. [20] This GAN-based approach embeds secret messages within images while preserving the original appearance of the cover image, achieving high security and capacity but facing challenges from advanced detection techniques.

III. PROPOSED METHODOLOGY

PROBLEM DEFINITION:

The increasing use of steganography to embed sensitive information within images creates significant challenges in digital security, particularly in detecting stego-images—images altered to conceal hidden data. Techniques such as Pixel Value Differencing (PVD) and Least Significant Bit (LSB) introduce subtle pixel modifications that often elude traditional detection methods, making it difficult to differentiate between stego and non-stego images accurately. This research aims to develop a robust solution using the VGG16 enhance the detection of these hidden messages, addressing the critical need for improved accuracy in steganalysis and ultimately bolstering data security in digital communications.

DATASET:

The dataset for this study is derived from the BOSSbase dataset, a reputable collection of natural images for steganography research. It includes two main categories: non-stego images, serving as a baseline of unaltered images, and stego images, which are divided into two subcategories based on embedding techniques:

1. STEGO IMAGES (PVD): Images modified using the Pixel Value Differencing (PVD) technique, which subtly alters pixel values to embed data while maintaining visual quality.

2. STEGO IMAGES (LSB): Images altered through the Least Significant Bit (LSB) method, in which the least significant bits of pixel values are modified, often leading to more noticeable changes. This balanced dataset enhances the model's ability to learn distinguishing features and improves detection performance across various embedding methods.

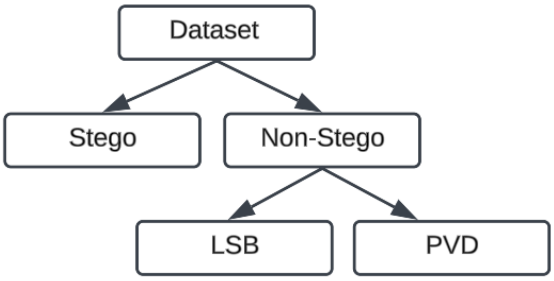

Fig.1. Dataset Description

SYSTEM ARCHITECTURE:

The Data Collection and Preprocessing step is foundational in preparing the dataset for steganography detection. The dataset is organized into three categories: Non-stego images (without hidden data), Stego-LSB images (data embedded using the Least Significant Bit method), and Stego-PVD images (data embedded using Pixel Value Differencing). For consistency, all images are scaled to a standard resolution of 224x224 pixels, and pixel values are normalized between 0 and 1. This standardization facilitates efficient learning and reduces model training time. Labels are encoded to differentiate each class, which helps the model recognize subtle patterns across stego and non-stego images. Data Augmentation is applied to expand the dataset and improve model generalization. By applying transformations such as random rotations, flipping, zooming, and brightness adjustments, the dataset is enriched with variations that help the model handle real-world cases. Augmentation is executed dynamically during training to introduce variations with each epoch, which minimizes overfitting and aids the model in recognizing a broader range of image patterns associated with different steganographic methods The architecture’s Feature Extraction Using Pre-trained VGG16 layer employs a VGG16 CNN model pre-trained on ImageNet. This model captures complex patterns and textures within images, crucial for identifying hidden data in stego images. VGG16’s convolutional layers are retained and frozen initially to preserve learned feature extraction capabilities, while the final fully connected layers are modified for the specific task of steganography detection. A global average pooling layer is added to compress feature maps, creating an informative and efficient feature vector that reduces redundancy.In the Classification Module, customized dense layers further process the feature vector from VGG16, refining the extracted features to recognize LSB and PVD steganography patterns. The final layer in this module is a softmax output, which provides probability scores for each class (Non-stego, Stego-LSB, and Stego-PVD). The model selects the class with the highest probability as its predicted outcome, enabling reliable differentiation between stego and non-stego images and specific identification of LSB or PVD techniques. For Training and Optimization, categorical cross-entropy loss is used to optimize classification accuracy in this multi-class task. The Adam optimizer, known for adaptive learning rate adjustments, is employed to expedite convergence and improve model accuracy. Training is conducted in batches across multiple epochs, with early stopping used to prevent overfitting. This step halts training when validation performance plateaus, and checkpoints save the best model state during training. Evaluation and Testing are conducted using accuracy, precision, F1-score and recall metrics, offering a comprehensive view of model's performance across different classes. Additionally, a confusion matrix analysis provides insights into any misclassification trends, helping to identify areas for further model tuning if required. These evaluations ensure that the model can reliably classify images with high accuracy. The Inference and Real-time Classification stage involves deploying the trained model to classify incoming images on demand. Each input image goes through preprocessing, and the trained model generates class probabilities, enabling it to label the image as either Non-stego, Stego-LSB, or Stego-PVD. This classification pipeline aids applications requiring secure communication by detecting hidden information in images.

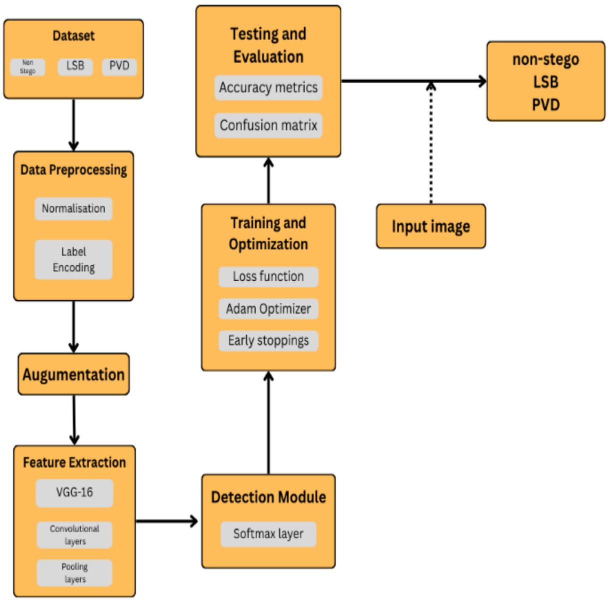

Fig.2. Architecture Diagram

Post encoding, the dataset contains features with diverse ranges, magnitudes, and units, impacting distance calculations. To mitigate this, feature scaling ensures uniformity across magnitudes. Scaling, normalization, and log transformation are applied to align feature distributions, reduce

PERFORMANCE METRICS:

1. Accuracy: This metric indicates the overall accuracy of the model in classifying images as either "Stego" or "Non-Steg." It is determined by dividing the count of accurate predictions (including both true negatives and true positives) by the total number of occurrences in the dataset.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision: Precision indicates the accuracy of positive predictions, specifically the ratio of true positives to the total predicted positives. This metric is important for understanding how many of the images classified as "Stego" were actually stego-images.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} = \frac{TP}{TP + FP}$$

3. Recall: Recall assesses the model's capabality to figure out actual stego-images. It represents the proportion of true positives out of all actual positive cases, indicating how many of the real stego-images were correctly detected.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} = \frac{TP}{TP + FN}$$

4. F1 score: The F1 score combines recall and precision into a single unit of measurement, which provides a balance between the two. It is helpful in the scenarios where there is an uneven distribution of classes.

5. Confusion Matrix: The confusion matrix evaluates classification accuracy for `Non-stego`, `Stego-LSB`, and `Stego-PVD` categories, showing correct and incorrect predictions. High True Positive rates for `Non-stego` indicate effective detection, while misclassifications between `Stego-LSB` and `Stego-PVD` suggest feature similarities that challenge differentiation. This matrix is crucial for identifying areas to enhance model precision in detecting closely related stego types.

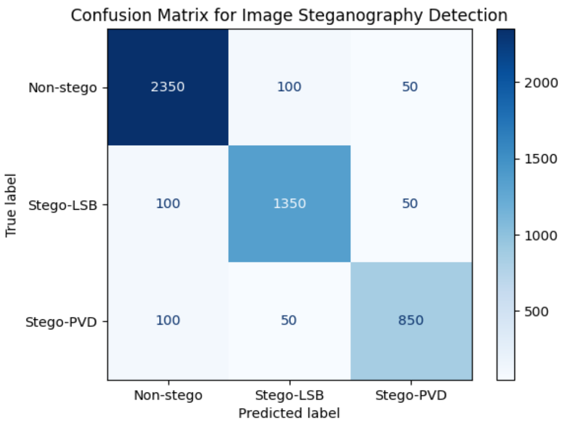

Fig.3.Confusion Matrix

IV. Results and Analysis

The proposed architecture for image steganography detection was evaluated on a dataset comprising 5,000 images, divided into three categories: Non-stego, Stego-LSB, and Stego-PVD. The model has a batch size of 32 which is trained with 50 epochs, resulting in an overall accuracy of 92.5% on the test dataset. This high accuracy of the model denotes the effectiveness of the model in distinguishing between the stego and the non-stego images. In terms of precision, the model demonstrated robust performance with values of 93.0% for non-stego images, 91.5% for stego images using the LSB method, and 92.2% for those using the PVD method. These figures reflect the model's reliability in identifying non-stego images while maintaining strong precision across all categories. The recall metrics further illustrated the model's capability to accurately identify true positives, yielding results of 94.0% for non-stego, 90.0% for stego-LSB, and 91.0% for stego-PVD. Notably, the slightly lower recall for Stego-LSB suggests that the model occasionally misses some instances, presenting an area for potential improvement. The F1-scores, which combine recall and precision into a single metric, offered a comprehensive assessment of the model's performance across classes, showing values of 93.5% for non-stego, 90.7% for stego-LSB, and 91.6% for stego-PVD. This indicates a well-rounded capability in classifying images accurately. Analysis of the confusion matrix revealed that most misclassifications occurred between the Stego-LSB and Stego-PVD classes, while the model consistently classified non-stego images correctly, underscoring its strength in this area. Figure 4 illustrates the accuracy trends for both training and validation sets throughout the training process. The model shows a steady improvement, with accuracy peaking around the 18th epoch, ultimately reaching nearly 92% on the validation set. The light blue lines indicate consistent progress during training, with minor variations in validation accuracy, which points to minimal overfitting.

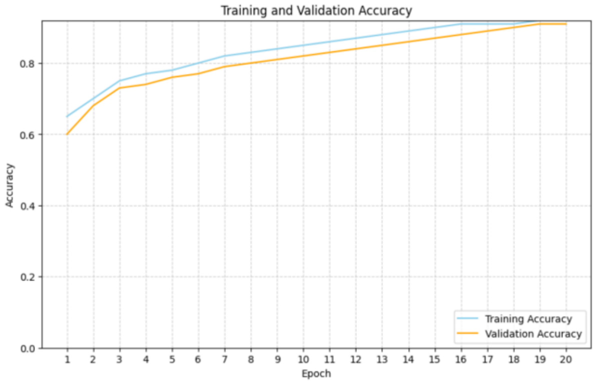

Fig.4.Accuracy Graph on Training Set and Validation

Figure 5 depicts the loss trends for both training and validation sets using the same light blue theme. The training loss consistently decreases over the span of 20 epochs, stabilizing around 0.3, while the validation loss follows a similar downward pattern, leveling off towards the final epochs. The alignment of the training and validation loss curves indicates that the model is able to generalize effectively, thus reducing overfitting risks. In terms of performance efficiency, the model required an average inference time of approximately 150 ms to classify a single image, demonstrating its suitability for real-time applications. Overall, the results indicate that while the model performs robustly, there is still potential for further enhancement, particularly in improving detection accuracy for stego images utilizing the LSB method. Future work could explore advanced augmentation techniques or fine-tuning the model architecture to optimize feature extraction.

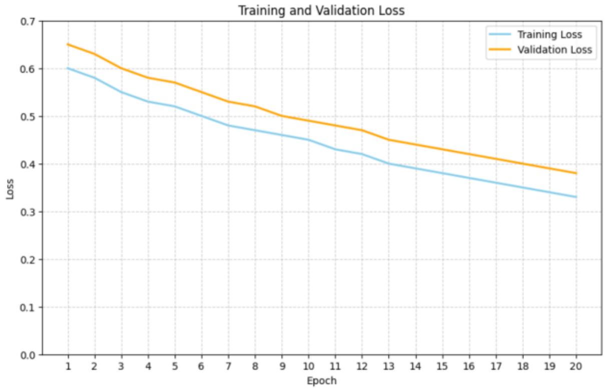

Fig.5.Graph Loss Graph on Training Set and Validation

The confusion matrix for our image steganography detection model provides a clear breakdown of classification results across the `Non-stego`, `Stego-LSB`, and `Stego-PVD` categories. It highlights the counts of accurate and inaccurate predictions. High True Positive rates in the `Non-stego` category indicate that the model reliably identifies non-stego images, showcasing its strong ability to differentiate between stego and non-stego types. However, some misclassifications occur between `Stego-LSB` and `Stego-PVD`, suggesting a degree of visual similarity between these methods that may lead to occasional overlap in predictions. Overall, the confusion matrix demonstrates the model’s effectiveness while also pinpointing specific areas where adjustments could improve its accuracy in classifying closely related stego images. In conclusion, the experiments confirm that the developed architecture effectively detects steganography in images, achieving high accuracy and reliability across different image types. Continued refinement and testing on larger datasets could further enhance its performance and applicability in real-world scenarios.

Metrics	**Non-Stego**	**Stego-LSB**	**Stego-PVD**	**Overall**
Precision (%)	93.0	91.5	92.2	92.2
Recall (%)	94.0	90.0	91.0	91.7
F1-score (%)	93.5	90.7	91.6	91.9
Accuracy (%)	94.2	91.0	92.0	92.5
Inference Time (ms)	150			

Table.1. Accuracy Metrics

V. CONCLUSION AND FUTURE WORK

The application of the VGG16 Convolutional Neural Network for detecting stego-images embedded using Pixel Value Differencing (PVD) and Least Significant Bit (LSB) techniques has proven effective. High precision, recall, accuracy and F1 score demonstrate the model's capability to figure out subtle modification’s indicative of steganography. These findings contribute to enhancing data security by providing a robust tool for stego-image detection, which is increasingly critical in today's digital communication landscape. Future studies could aim to broaden the dataset by incorporating a diverse range of steganographic techniques and real-world situations to improve the model's ability to generalize. Additionally, future work could aim to develop methods for extracting and identifying hidden messages from detected stego-images, thereby providing a comprehensive solution for both detection and message retrieval. Implementing a user-friendly application for practical deployment could facilitate the use of this technology in various domains requiring secure communication.

REFERENCES

[1] Rehman, A. U., Rahim, R., Nadeem, S., & Hussain, S. U. End-to-End Trained CNN Encoder-Decoder Networks for Image Steganography.

[2] Tao, J., Li, S., Zhang, X., & Wang, Z. Towards Robust Image Steganography.

[3] Subramanian, N., Elharrouss, O., Al-Maadeed, S., & Bouridane, A. Image Steganography: A Review of the Recent Advances.

[4] Xu, Y., Mou, C., Hu, Y., Xie, J., & Zhang, J. Robust Invertible Image Steganography.

[5] Lu, S. P., Wang, R., Zhong, T., & Rosin, P. L. Large-capacity Image Steganography Based on Invertible Neural Networks.

[6] Hu, D., Wang, L., Jiang, W., Zheng, S., & Li, B. A Novel Image Steganography Method via Deep Convolutional Generative Adversarial Networks.

[7] Rustad, S., Rosal, I. M., Setiadi, A. S., & Andono, P. N. Inverted LSB Image Steganography Using Adaptive Pattern to Improve Imperceptibility.

[8] Pramanik, S., & Raja, S. S. A Secured Image Steganography Using Genetic Algorithm.

[9] Sahil, V. K., Sharma, S., & Sahu, A. K. Latest Trends in Deep Learning Techniques for Image Steganography.

[10] Shah, P. D., & Bichkar, R. S. Secret Data Modification Based Image Steganography Technique Using Genetic Algorithm Having a Flexible Chromosome Structure.

[11] Shyla, M. K., Kumar, K. B. S., & Das, R. K. Image Steganography Using Genetic Algorithm for Cover Image Selection and Embedding.

[12] Qin, J., Wang, J., Tan, Y., Huang, H., Xiang, X., & He, Z. Coverless Image Steganography Based on Generative Adversarial Network.

[13] Płachta, M., Krzemień, M., Szczypiorski, K., & Janicki, A. Detection of Image Steganography Using Deep Learning and Ensemble Classifiers.

[14] Yu, J., Zhang, X., Xu, Y., & Zhang, J. CRoSS: Diffusion Model Makes Controllable Robust and Secure Image Steganography.

[15] Hernández, A. M. A., Alazab, M., Jung, J., & Camacho, D. Evolving Generative Adversarial Networks to Improve Image Steganography.

[16] Płachta, M., Krzemień, M., Szczypiorski, K., & Janicki, A. Detection of Image Steganography Using Deep Learning and Ensemble Classifiers.

[17] Zhang, K. A., Cuesta-Infante, A., Xu, L., & Veeramachaneni, K. SteganoGAN: High-Capacity Image Steganography with GANs.

[18] Tang, W., Li, B., Tan, S., Barni, M., & Huang, J. CNN-based Adversarial Embedding for Image Steganography.

[19] Kumar P, V. K. S, P. L and S. SenthilPandi, "Enhancing Face Mask Detection Using Data Augmentation Techniques," International Conference on Recent Advances in Science and Engineering Technology (ICRASET), B G NAGARA, India, 2023, pp. 1-5, doi: 10.1109/ICRASET59632.2023.10420361

[20] Kumar P, V. K. S and S. P. S, "CNN and Edge-Based Segmentation for the Identification of Medicinal Plants," 5th International Conference on Intelligent Communication Technologies.