*A project report on*

# EXPLAINABLE STEGANALYSIS FRAMEWORK USING CSRNET WITH ADAPTIVE SRM FILTERING AND INTEGRATION OF XAI

*Submitted in partial fulfillment for the award of the degree of*

## M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics

*by*

**SAKTHI SAIRAM.U (21MIA1087)**



**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

November, 2025

# EXPLAINABLE STEGANALYSIS FRAMEWORK USING CSRNET WITH ADAPTIVE SRM FILTERING AND INTEGRATION OF XAI

*Submitted in partial fulfillment for the award of the degree of*

**M.Tech. (Integrated) Computer Science and Engineering with Specialization in Business Analytics**

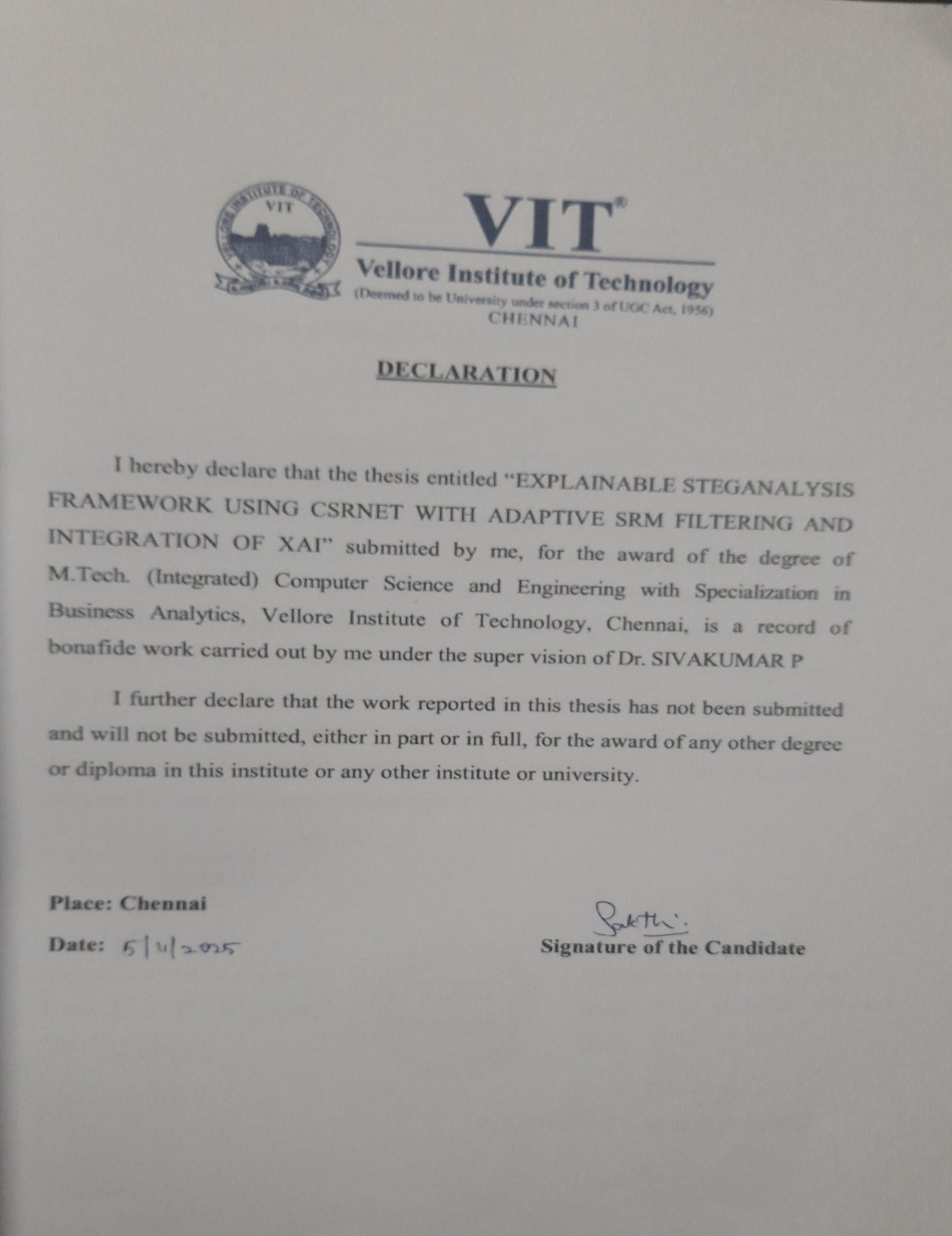
*by*

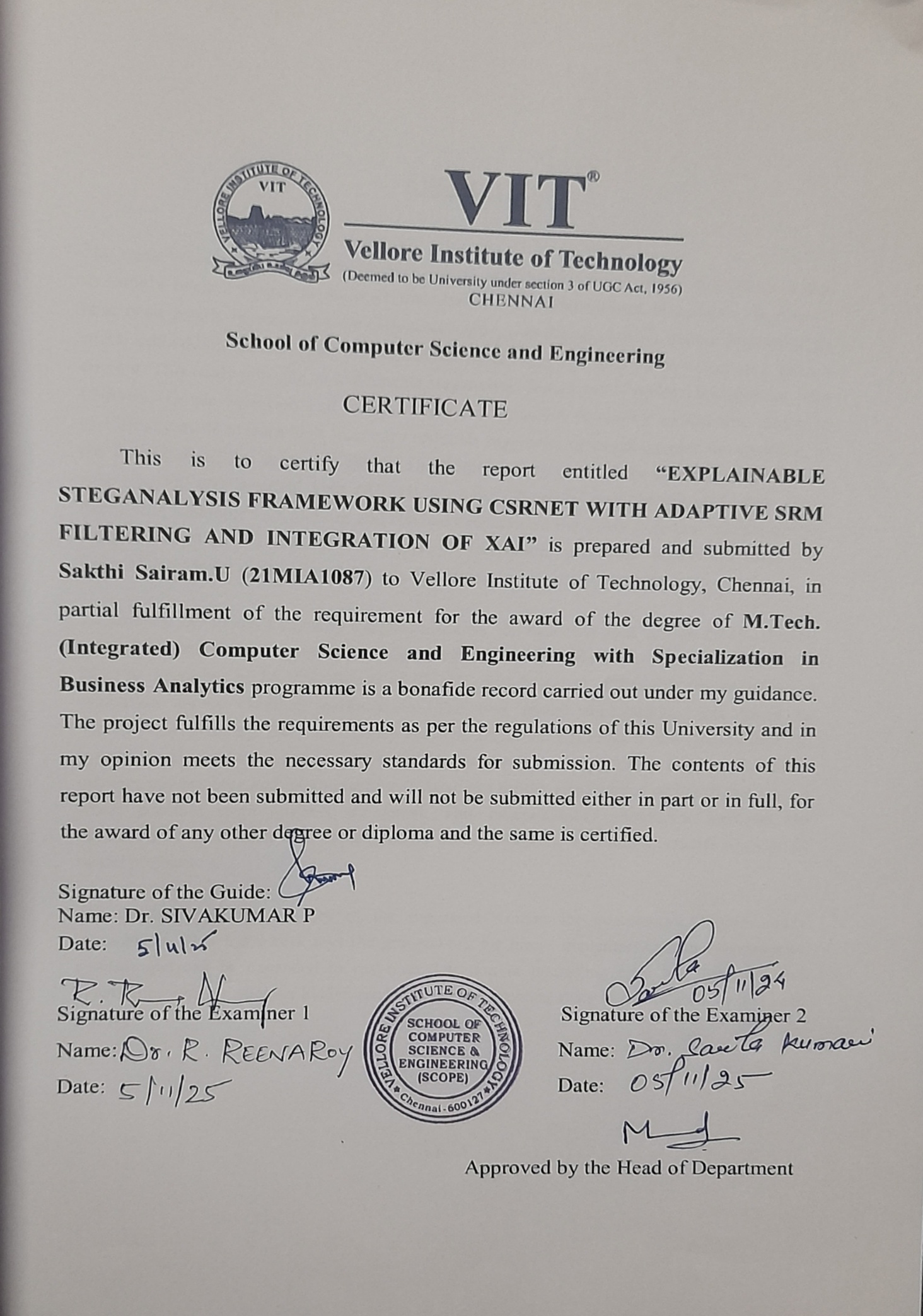
**SAKTHI SAIRAM.U (21MIA1087)**



**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

November, 2025





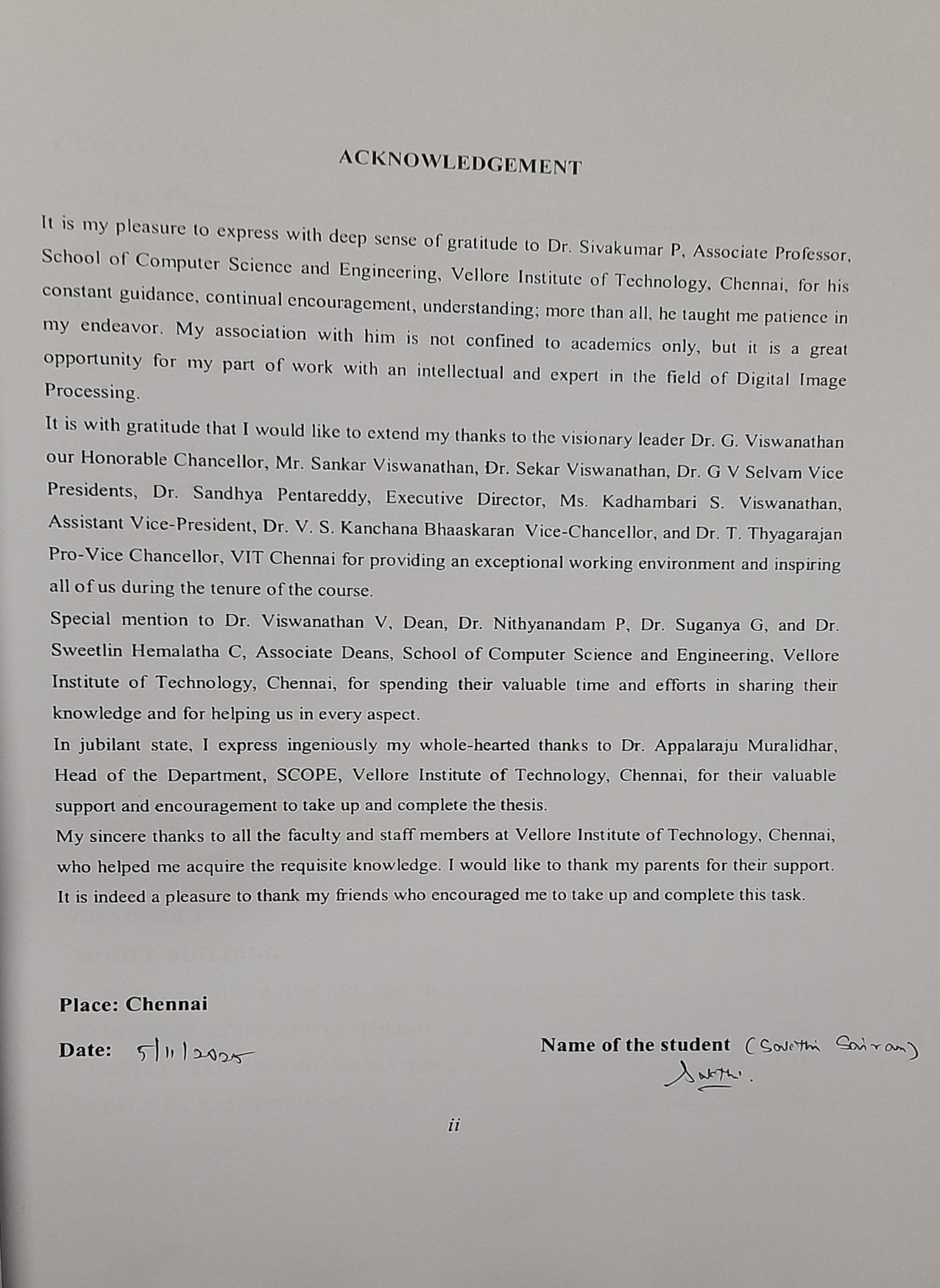
**ABSTRACT**

The risks associated with steganography, where adversaries embed secret information within seemingly benign images to facilitate unauthorized data transmission or malware distribution have been greatly increased by the exponential growth in digital communication. Because of its subtle embedding patterns that imitate natural noise, it is still difficult to detect such hidden content. This study presents an explainable steganalysis framework to address this problem. The CSRNet (Cover Selection and Adaptive Filtered Residual Network) architecture, which is specifically made to learn high-frequency residuals suggestive of steganographic manipulation, is the foundation of the suggested model. The framework combines three essential elements: the Adaptive Filtered Residual Unit (AFRU) for gated attention-based feature refinement using depth-wise separable convolutions and residual connections, the Cover Selection Module (CSM) for patch-level region identification based on temperature-controlled importance maps and Spatial Rich Model (SRM) filter banks for improving residual noise patterns.

To increase the model’s capacity for generalization, the dataset is split into an 80-10-10 train-test-validation split and extensive data augmentation techniques are used. With Test Accuracy of 88.21%, Precision of 86.35%, Recall of 87.13%, and F1-score of 86.16%, the experimental evaluation shows excellent performance. The best model configuration was found by hyperparameter tuning. DeepLIFT (Deep Learning Important Features), a XAI (Explainable AI) is used to calculate mean and top attribution scores in relation to a CSRNet baseline in order to increase transparency and interpretability. This allows for the visualization of significant regions where possible stego artifacts are found. This interpretability component increases the dependability of forensic and security-based decision-making by offering crucial support for model predictions.

The results verify the efficacy of the suggested method for interpretable, high-accuracy steganalysis. This framework lays the groundwork for future implementation in cybersecurity systems and can be expanded to real-time detection in messaging apps like Telegram and WhatsApp, where malicious hidden payloads present growing risks to device security and user privacy.

*i*



## CONTENTS

CONTENTS iii

LIST OF FIGURES iv

LIST OF TABLES v

LIST OF ACRONYMS vi

CHAPTER 1

INTRODUCTION

* 1. INTRODUCTION 1
  2. SCOPE OF THE PROJECT 2
  3. OBJECTIVES 2
  4. PROBLEM STATEMENT 3
  5. RELATED WORKS 4

CHAPTER 2

DATASET DESCRIPTION

* 1. OVERVIEW OF THE DATASET 6
  2. VISUALIZATION AND INSIGHTS 7

CHAPTER 3

MODEL BUILDING

* 1. DATA AUGMENTATION AND PRE-PROCESSING 11
  2. MODEL BUILDING AND COMPILING 12
  3. MODEL TRAINING AND VALIDATION 14
  4. DEEPLIFT XAI IMPLEMENTATION 16

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

* 1. PERFORMANCE OF CSRNET MODEL 18
  2. PERFORMANCE OF DEEPLIFT XAI 20

CHAPTER 5

CONCLUSION AND FUTURE WORK

* 1. CONCLUSION 23
  2. FUTURE WORK 24
  3. CHALLENGES FACED 25
  4. LIMITATIONS 26

**APPENDICES**

APPENDIX A DATASET INFORMATION 28

APPENDIX B HYPERPARAMETER TUNING 28

APPENDIX C MODEL ARCHITECTURE DETAILS 29

APPENDIX D EXPERIMENT ENVIRONMENT 30

APPENDIX E EVALUATION DETAILS 30

APPENDIX F DEEPLIFT CONFIGURATION 31

CODE SNIPPETS

1. DATASET IMPORTATION 32
2. PGM TO PNG CONVERSION 32
3. DATA LABELING AND PRE-PROCESSING 33
4. COVER SELECTION MODULE (CSM) 34
5. SRM FILTERBANKS 35
6. ADAPTIVE FILTERED RESIDUAL BLOCK (AFRU) 36
7. CSRNET 37
8. CSRNET SUMMARY 38
9. DATA TRANSFORMATION AND SPLITTING 39
10. MODEL TRAINING 40
11. MODEL EVALUATION 41
12. MODEL TESTING 41

REFERENCES

*iii*

### LIST OF FIGURES

* 1. PIXEL INTENSITY HISTOGRAM OF SAMPLE COVER IMAGES 8
  2. PIXEL INTENSITY HISTOGRAM OF SAMPLE STEGO IMAGES 8
  3. MEAN HEATMAP OF SAMPLE COVER IMAGES 9
  4. MEAN HEATMAP OF SAMPLE STEGO IMAGES 10
  5. CSRNET MODEL ARCHITECTURE 13
  6. CSRNET MODEL WORKFLOW 14
  7. DEEPLIFT ATTRIBUTION HEATMAP FOR STEGO IMAGES 20
  8. DEEPLIFT ATTRIBUTION HEATMAP FOR COVER IMAGES 21

*iv*

**LIST OF TABLES**

* 1. RESULT METRICS 18
  2. HYPERPARAMETER TUNING RESULTS 19

APPENDIX 1 HYPERPARAMETER CONFIGURATION 28

APPENDIX 2 EXPERIMENT ENVIRONMENT DETAILS 30

*v*

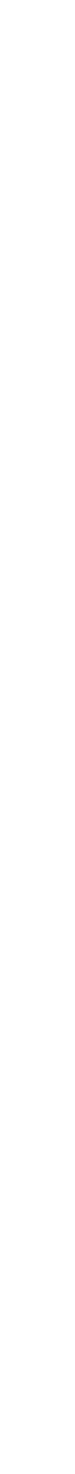
**LIST OF ACRONYMS**

AFRUWAP Adaptive Filtered Residual Unit

API Application Programming Interface

BOSSBase Break Our Steganographic System Base

CSM Cover Selection Module

CSRNet Cover Selection and Adaptive Filtered Residual Networks

DeepLIFT Deep Learning Important Features

GPU Graphic Processing Unit

LSB Least Significant Bit

LRP Layer-wise Relevance Propagation

SPAM Subtractive Pixel Adjacency Matrix

SRM Spatial Rick Model

PGM Portable Gray Map

ReLU Rectified Linear Unit

RGB Red Green Blue

XAI Expalainable Artificial Intelligence

*vi*

**Chapter 1**

**INTRODUCTION**

* 1. INTRODUCTION

The need for secure information exchange has grown in importance due to the exponential growth of digital communication. Although cryptography uses encryption to guarantee confidentiality, the presence of encrypted data itself frequently draws malevolent attention. On the other hand, steganography uses hidden data embedded in multimedia content like images, audio, or video to hide communication itself. Although this method is frequently employed for legitimate purposes, it has also emerged as a potential threat vector for cybercriminals who use it to spread malware, carry out illegal communications, or launch covert attacks. As a result, steganalysis, the field of identifying and deciphering hidden data has become increasingly important in the fields of digital forensics and cybersecurity. Conventional steganalysis methods mainly rely on statistical patterns and hand-crafted features, which frequently don’t generalize across different steganographic algorithms and embedding payloads. Since deep learning-based frameworks can automatically extract subtle and high-level spatial features, they have become powerful alternatives to overcome these limitations. Specifically, CSRNet has proven to be very effective at capturing variations at the pixel level caused by steganographic embedding. Furthermore, by highlighting noise residuals and artifact patterns, incorporating adaptive filtering techniques like AFRU, CSM and SRM improves feature representation even more.

However, since deep learning-based steganalysis systems are difficult to understand, they are frequently referred to as “black-box” models, which presents significant difficulties in regulated fields like forensics and cybersecurity. DeepLIFT is used to address this problem by offering significant attribution maps that support the model’s conclusions. In addition to increasing transparency and trust, this helps analysts identify which pixel regions or residual patterns affected classification results. In order to provide precise, comprehensible, and trustworthy detection of hidden data within images, this project suggests an Explainable Steganalysis Framework that combines CSRNet with adaptive SRM-based filtering and integrates DeepLIFT XAI.

* 1. SCOPE OF THE PROJECT

The design, development, assessment, and interpretation of an explainable deep learning based steganalysis framework that accurately finds hidden information embedded in digital images are all included in the project’s scope. By incorporating interpretability mechanisms, this system ensures its applicability in real-world cybersecurity and forensic environments, going beyond simple classification accuracy. By adding CSM, SRM filter banks, and AFRU, which together amplify pixel distortions commonly caused by steganographic payloads, the framework further improves performance. DeepLIFT calculates the mean and top feature attributions for specific test images using the trained CSRNet architecture as the model’s baseline. In real-world situations, the explanation mechanism improves interpretability, trust, and traceability by highlighting pixel regions and residual signatures that were most important in the classification decision.

In addition to XAI-driven interpretability insights, standard evaluation metrics like Accuracy, Precision, Recall, and F1-score are used to gauge the framework’s performance. To maximize learning performance, hyperparameter tuning experiments are carried out, and the outcomes are recorded in a comparative table. The project’s scope also includes possible deployment scenarios, especially in cybersecurity monitoring systems, regulatory compliance tools, and digital forensic pipelines. Additionally, it establishes the groundwork for future modifications that could identify malicious steganography shared through instant messaging services (like Telegram and WhatsApp), where attackers might conceal malware or private information in pictures without the user’s knowledge.

* 1. OBJECTIVES

Designing and implementing an explainable steganalysis framework that successfully identifies hidden steganographic content in grayscale images while offering clear, comprehensible explanations for each prediction is the main objective of this research. The study focuses on using the CSRNet architecture enhanced with attention guided residual learning to accomplish this. The following are the specific goals:

* To create a CSRNet-based steganalysis model that effectively captures subtle artifacts and high-frequency residuals brought about by steganographic embedding methods.
* Using temperature-based attention mapping, the CSM is integrated to dynamically identify and localize image regions that are likely to contain embedded stego signals.
* To use SRM filter banks, such as Laplacian, Laplacian-8, High-Pass, Sobel X, and Sobel Y filters, to improve the extraction of residuals that resemble noise and suggest the existence of hidden data.
* To enhance feature sensitivity and learning efficiency by implementing the AFRU for selective feature refinement using gated attention, residual connections, and depth-wise separable convolutions.
* To assess the efficacy of the suggested model on the BOSSBase (Break Our Steganographic System Base) dataset with balanced cover and stego distributions using common performance metrics like Accuracy, Precision, Recall, and F1-score.
* By producing attribution maps that emphasize decision-critical image regions and offering insightful justification for the model’s classification results, DeepLIFT XAI will improve interpretability.
* To determine whether the framework can be implemented in actual cybersecurity systems, especially for proactive steganography detection in digital communication platforms.
  1. PROBLEM STATEMENT

Steganography has become a sophisticated technique for hiding sensitive or malicious data within innocuous-looking images due to the quick expansion of digital communication and social media platforms. Cybercriminals are increasingly using these methods to enable covert data exfiltration, malware distribution, and unauthorized information sharing, even though they can be used for legitimate purposes like digital watermarking or secure communication. When stego changes are subtle and visually indistinguishable, traditional steganalysis methods based on hand-crafted statistical features and shallow classifiers are not robust enough to identify contemporary adaptive embedding techniques.

By automatically discovering hidden feature patterns, deep learning has improved steganalysis performance. Nevertheless, a significant drawback still exists: the majority of deep models function as “black boxes” offering no explanations or insights into their predictions. Trust is damaged by this lack of interpretability, particularly in cybersecurity and forensic investigations where explanation-driven evidence is essential. Furthermore, in situations with low payloads or highly adaptive stego schemes, deep models frequently neglect to concentrate on the precise areas where steganographic changes take place, decreasing detection accuracy. Thus, a steganalysis framework that not only accurately detects embedded stego content but also offers clear, understandable explanations that show how and where the model detects possible tampering is desperately needed. In order to close this gap, this study suggests an explainable CSRNet-based steganalysis model that is improved with SRM filtering and DeepLIFT XAI. It is intended to increase detection accuracy while guaranteeing interpretability for practical application in cybersecurity and forensic settings.

* 1. RELATED WORKS

For more than 20 years, steganalysis has been a dynamic field of study, progressing from manually created feature-based models to deep learning-driven and now XAI enabled frameworks. Finding anomalies in pixel distributions or statistical inconsistencies introduced during data embedding was the main goal of early steganalysis techniques. In order to identify minute embedding artifacts, conventional methods like SRM and SPAM (Subtractive Pixel Adjacency Matrix) extracted high-dimensional handcrafted features from residual maps. These approaches worked well for straightforward algorithms like LSB (Least Significant Bit) steganography, but they had trouble with adaptive or high-capacity embedding methods that strategically distribute changes throughout the image. The performance of steganalysis was greatly improved with the introduction of deep learning. One of the earliest CNN-based steganalysis models, called Xu-Net, was presented by Xu et al. (2016). It automatically extracted discriminative spatial features from images without the need for manual preprocessing. Ye-Net (2017) expanded on this concept by adding batch normalization and SRM-inspired fixed high-pass filters, which enhanced robustness and convergence. By combining SRM filters with convolutional blocks, Yedroudj-Net (2018) further refined this architecture and achieved greater accuracy on the BOSSBase and BOWS2 datasets. The shift from manual feature engineering to data-driven feature extraction was signaled by these models.

Despite their achievements, the majority of deep learning-based steganalysis networks functioned as black-box classifiers, providing high accuracy but little interpretability. In order to improve the detection of subtle embedding traces while maintaining model depth efficiency, researchers started investigating architectures such as SRNet (2019), which employed residual learning. An additional advancement was the CSRNet architecture, which uses adaptive residual filtering and cover selection to concentrate on areas of the image most likely impacted by steganography. This flexibility decreased false positives and increased feature sensitivity. XAI has gained popularity to solve the interpretability issue in deep learning in tandem with model advancements. By emphasizing which pixels or regions affect classification decisions, techniques like Grad-CAM, LIME, and DeepLIFT offer visual explanations. Although XAI has been extensively employed in security, sentiment analysis, and medical imaging, its use in steganalysis is still quite limited. Few studies have directly incorporated XAI into the model evaluation process, despite recent research exploring explainability to understand how CNNs detect embedding noise or texture artifacts. Thus, by combining CSRNet with adaptive SRM filtering and DeepLIFT explainability, this project expands upon these foundations and creates a cohesive framework that enhances detection accuracy while also offering comprehensible visual insights. By bridging the gap between high-performance detection and model transparency, the suggested system enables researchers and practitioners to comprehend why an image is categorized as stego or cover. This contribution presents the work as a significant step toward reliable and comprehensible steganalysis systems appropriate for practical cybersecurity and digital forensics applications.

**Chapter 2**

**DATASET DESCRIPTION**

* 1. OVERVIEW OF THE DATASET

The BOSSBase1.01 dataset, a well-known benchmark for image steganalysis research, is used in this study. The dataset, which was created by the BOSS community, includes 20,000 uncompressed grayscale photos that were initially saved in Portable Gray Map (PGM) format. Several camera devices contributed to the natural photographic scenes, which captured a wide variety of textures, lighting, and environmental changes. To ensure consistency for feature extraction and model training in deep learning-based steganalysis architectures, each image is standardized to a resolution of 512×512 pixels. Two classes are equally assigned to the dataset:

* Cover photos (10,000 samples): These are pure, unaltered grayscale photos with no hidden data embedded in them.
* Stego images (10,000 samples): These images have embedded payloads that were created using different spatial domain steganographic algorithms (such as HUGO, WOW, and S-UNIWARD). These payloads introduce minute pixel-level distortions that are statistically comparable to natural noise, making detection extremely difficult.

The model is trained without class bias thanks to the balanced distribution of cover and stego samples, which facilitates accurate learning of discriminative features. The model can concentrate on texture irregularities and high-frequency residuals instead of being affected by color noise or RGB channel variations because all images are grayscale. Because of this, BOSSBase is especially well-suited for deep steganalysis tasks that depend on pixel residuals and edge-level artifacts.

Additionally, a wide variety of statistical noise patterns are introduced by the dataset’s diverse content, which includes landscapes, human subjects, urban settings, foliage, and indoor scenes. This variation improves sensitivity to minute distortions introduced during steganographic processes and strengthens the model’s capacity to generalize across various embedding contexts. When assessing contemporary architectures such as CSRNet, which rely on residual-aware learning and attention-driven localization to detect hidden payloads with high accuracy, the use of BOSSBase is particularly important. The dataset offers significant visual areas where DeepLIFT Explainable AI can produce attribution-based explanations in order to facilitate model interpretability. Because of this compatibility, BOSSBase is a great option for explainability-focused steganalysis as well as high-performing detection.

* 1. VISUALIZATION AND INSIGHTS

To ensure meaningful feature extraction and well-informed architectural decisions, a thorough understanding of the dataset is essential before beginning any preprocessing or model development. In order to uncover the underlying features of both cover and stego images, data visualization is essential. This is achieved by choosing suitable charts, analyzing distributional patterns, and spotting anomalies. Pixel intensity distributions across grayscale images were examined in this study using histogram-based visualization, which provided insights into the statistical behavior of embedded versus unaltered content. The cover image histograms (Fig. 1) show smooth, continuous distributions consistent with typical features of natural images. These distributions show how different lighting conditions, object textures, contrast levels, and environmental complexity all contribute to intrinsic variations in image scenes. A baseline reference for the pixel distribution of an unaltered image is established by such natural variability.

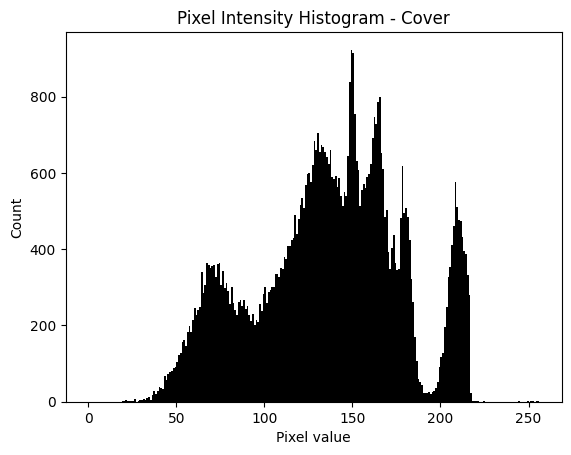
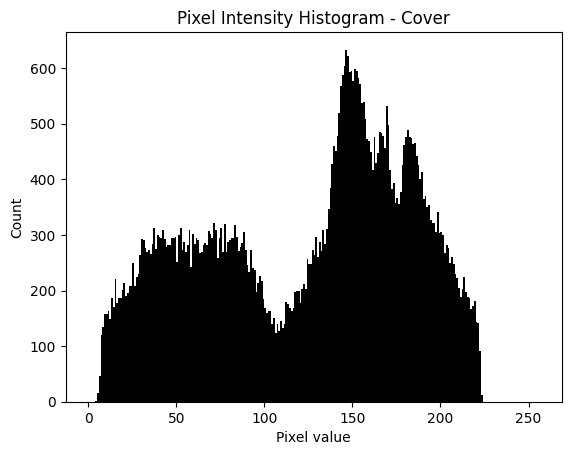


Fig.1 (Pixel Intensity Histogram of sample Cover images)

As an illustration of the systematic changes brought about by steganographic embedding, the histograms of stego images (Fig. 2) show clear distortions in the form of sharp peaks, bi-modal structures, and fluctuating intensity clusters. The influence of LSB steganography, a popular embedding technique that slightly modifies pixel values without noticeably altering the image, is suggested by the presence of prominent spikes at intensity extremes, especially close to pixel value 255.

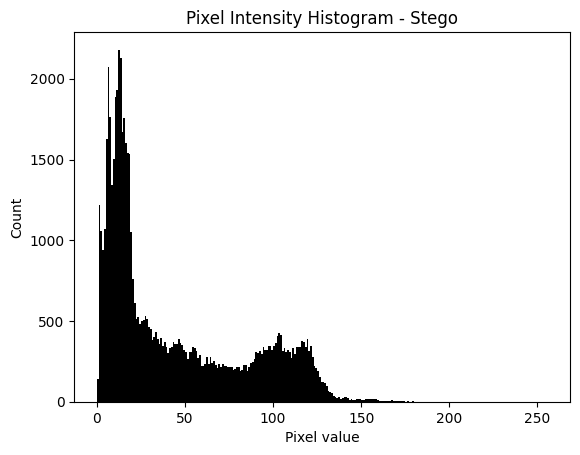
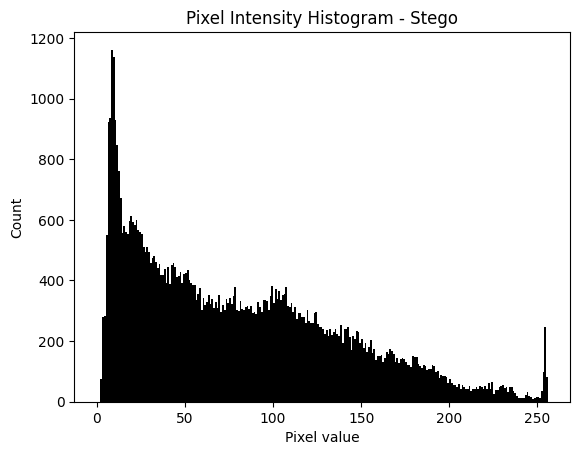


Fig.2 (Pixel Intensity Histogram of sample Stego images)

The use of adaptive steganography techniques, which strategically concentrate data embedding in textured or high-noise areas to evade detection, is also highlighted by irregular peaks and valleys. The significance of employing feature-sensitive models like CSRNet, which extract residual-based patterns and highlight regions probably impacted by embedding, is reinforced by these histogram anomalies, which act as early warning signs of stego artifacts. In general, visualization directs the development of residual-aware filters, patch selection mechanisms, and interpretability strategies in the suggested framework in addition to statistically differentiating cover and stego images.

To visually examine pixel-level variations across several cover and stego samples, mean intensity heatmaps were created in addition to histogram analysis. The cover image heatmaps (Fig. 3) show gradients that flow smoothly and naturally, reflecting the coherent spatial structures found in unaltered grayscale photos. These heatmaps show distinct spatial features like object boundaries, horizon lines, and prominent subjects encircled by structured backgrounds by maintaining constant transitions between pixel intensities. The lack of random patches or sudden geometric breaks supports the cover images’ innate consistency and indicates that their spatial regularity has not been disturbed by artificial embedding.

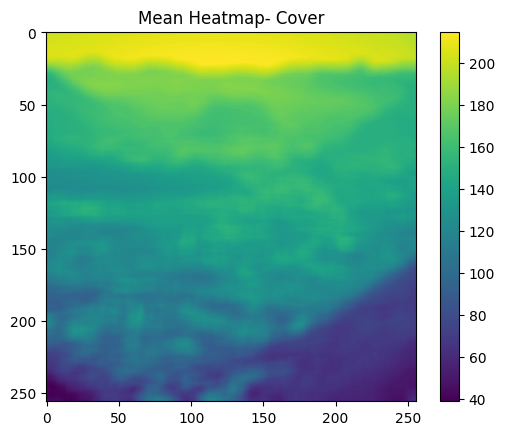
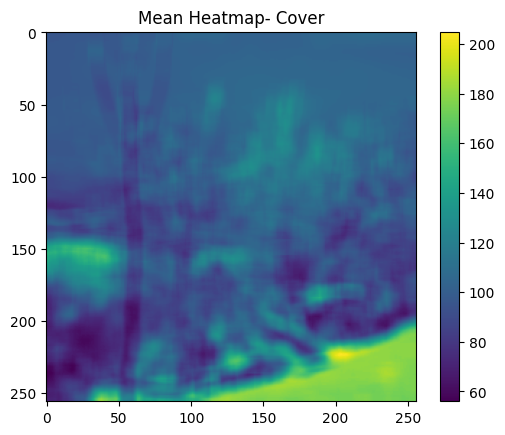


Fig.3 (Mean Heatmap of sample Cover images)

On the other hand, significant spatial inconsistencies and artifact-like areas that diverge from natural image distributions are visible in the heatmaps produced from stego images (Fig. 4). These visualizations frequently show localized distortions that are frequently linked to steganographic embedding, such as blocky areas, abrupt discontinuities, and asymmetrical geometric shapes. These artifacts result from changes made during payload insertion, particularly in adaptive steganography techniques that minimize visual detectability by selectively altering higher-textured and noise-prone regions. The heatmaps’ uneven distribution of high-intensity zones is a crucial sign of hidden data influence, indicating that mean heatmaps can successfully draw attention to stego-induced residual deviations.

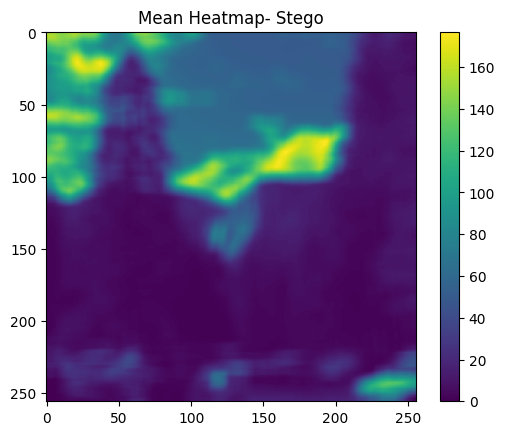
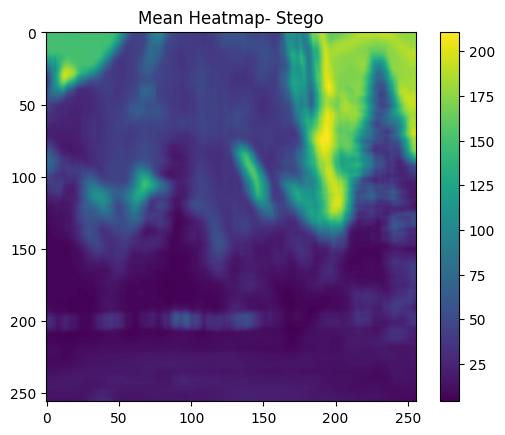


Fig.4 (Mean Heatmap of sample Stego images)

All things considered, these visualizations highlight an important realization. Although statistical and spatial patterns in cover and stego images may show minute but discernible differences, these variations are usually too subtle to be reliably identified through manual human inspection. This emphasizes the need to use sophisticated deep learning models, like CSRNet, which are made especially to capture spatial discrepancies and fine-grained residual cues resulting from steganographic embedding. Visualization methods like heatmaps and histograms provide a crucial basis for confirming the significance of automated, residual-aware steganalytic frameworks by methodically exposing both natural consistency and artificial disruptions.

**Chapter 3**

**MODEL BUILDING**

* 1. DATA AUGMENTATION AND PRE-PROCESSING

A thorough data preprocessing and augmentation pipeline was created in order to increase the dataset’s variability, boost generalization, and guarantee reliable model learning prior to training the suggested CSRNet-based steganalysis framework. Preprocessing was necessary to standardize image formats and get the grayscale, PGM-formatted images in the BOSSBase dataset ready for deep learning workflows. To enable supervised learning and classification, each image was labeled as either cover (0) or stego (1) and transformed into a compatible format. During training and assessment, label encoding made sure the model could discriminate between naturally occurring (unaltered) and steganographically embedded images. To address the dataset’s limited diversity of spatial textures and structural patterns, data augmentation was done. To mimic changes in orientation, illumination, and scale that the model might experience in real-world applications, augmentation techniques like random resizing, flipping, rotation, and brightness adjustment were used. The dataset’s variability was further enhanced by applying random contrast, saturation, and hue adjustments, which made the network resistant to small visual distortions and inconsistent lighting. The CSRNet model learned generalized features instead of memorizing particular pixel distributions thanks to these transformations, which were crucial in preventing overfitting.

The augmented dataset was divided into subsets for training (80%), validation (10%), and testing (10%), with the same percentage of cover and stego images in each. Consistent evaluation and balanced learning were guaranteed by this stratified split. To stabilize the training process and speed up convergence, all images were normalized to a fixed range (usually [0,1]) before being fed into the network. Maintaining high image fidelity during preprocessing was crucial because steganographic artifacts frequently appear as tiny variations in pixel intensity. Because they might unintentionally remove important residual features necessary for steganalysis, no aggressive noise reduction or smoothing operations were carried out. The dataset was successfully diversified while maintaining the subtle steganographic cues required for precise detection thanks to this methodical augmentation and preprocessing pipeline. The CSRNet model’s capacity to identify embedded content under a variety of image conditions and steganographic embedding techniques was enhanced by this meticulous preparation, which allowed it to concentrate on significant residual information.

* 1. MODEL BUILDING AND COMPILIING

The CSRNet architecture, which was created especially to identify minute steganographic artifacts while preserving interpretability, is the foundation of the suggested Explainable Steganalysis Framework (Fig. 5). The CSM, SRM Filter Banks, and the AFRU are the three main parts of the model that work together to systematically capture, refine, and analyze noise-like residuals that are suggestive of hidden data embedding. The first layer of the architecture, the CSM, is in charge of identifying possible areas of an image where steganographic changes are most likely to take place. It uses a temperature-controlled sigmoid function to create an importance map (A) after segmenting the image into patches. The softness of selection is determined by the temperature parameter; higher values result in smoother, probabilistic attention maps, while lower values produce sharper, binary-like maps. Instead of processing the entire image uniformly, the CSM uses three convolutional layers that gradually extract spatial dependencies and produce attention maps that highlight embedding-prone areas. This allows the model to concentrate its learning on the most informative regions.

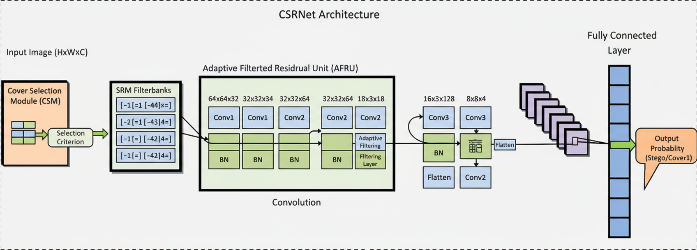


Fig.5 (CSRNet Model Architecture)

After CSM, the SRM filter bank layer amplifies high-frequency components and suppresses irrelevant content using five different high-pass filters: Laplacian, Laplacian-8, High-Pass (HP), Sobel X, and Sobel Y. By highlighting edge patterns and residual noise structures that are frequently undetectable in the spatial domain, these filters amplify minute pixel-level anomalies that result from steganographic embedding. After filtering, the outputs are fed into the AFRU for adaptive gating and feature refinement. The AFRU employs batch normalization for stability, dropout regularization to avoid overfitting, depth-wise separable convolutions for computational efficiency, pointwise convolutions for feature fusion, and ReLU activation for non-linearity. While the gating mechanism multiplies feature maps with attention weights derived from CSM, highlighting areas likely to contain stego alterations, a skip connection guarantees gradient flow and stabilizes training.

These modules’ outputs are combined and flattened into fully connected layers for classification by the integrated architecture. Using a sigmoid activation function, the last layer generates binary predictions: 0 for cover images and 1 for stego images. The model employs the binary cross-entropy loss function optimized with the Adam optimizer during compilation, which is renowned for its effective gradient handling and adaptive learning rate. To balance model stability and convergence speed, the learning rate was experimentally adjusted. To thoroughly assess model performance during training and validation, performance metrics including Accuracy, Precision, Recall, and F1-score were incorporated. Using an 80:10:10 train-validation-test split and GPU acceleration for effective computation, the CSRNet model was trained using the prepared dataset. To avoid overfitting and guarantee ideal convergence, frequent checkpoints and early stopping were used. The combined CSRNet model produced a reliable and comprehensible steganalysis pipeline that could identify even the most minute embedding traces in grayscale images by combining spatial filtering, adaptive attention, and residual learning.

* 1. MODEL TRAINING AND VALIDATION

To guarantee the robustness, generalization, and interpretability of the model, the Explainable Steganalysis Framework’s training and validation phase using CSRNet with Adaptive SRM Filtering and XAI Integration was carefully planned (Fig. 6). Following the construction of the CSRNet architecture, the model underwent a thorough training regimen designed to prevent overfitting to irrelevant noise patterns while successfully learning subtle spatial perturbations introduced by steganographic embedding techniques. This dataset split preserved unseen samples for objective assessment while ensuring the model had enough information to learn intricate patterns. To ensure consistency in input dimensions and to maximize GPU memory utilization during training, all images were resized to a fixed dimension, usually 256x256 pixels. To improve convergence and stabilize gradient updates, the pixel values were normalized within the interval [0,1].

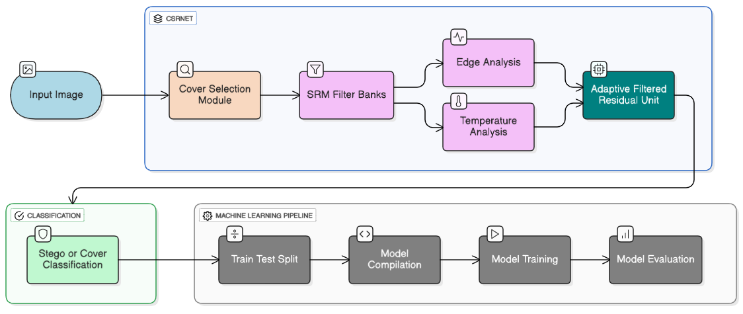


Fig.6 (CSRNet Model Workflow)

The model used the binary cross-entropy loss function to measure the difference between the actual and predicted labels during the training phase. Since the Adam optimizer adaptively modifies learning rates for each parameter, guaranteeing effective and stable convergence, it was employed with an initial learning rate of 1e-4. After multiple epochs of stagnant validation accuracy, a learning rate scheduler was incorporated to lower the rate by a factor of 0.1, allowing for optimization process fine-tuning in later training stages. Data augmentation techniques were integrated into the training pipeline to improve the network’s capacity for generalization. To simulate various conditions and reduce overfitting, these included random rotations, horizontal and vertical flips, Gaussian noise injection, and brightness variations. To further inhibit neuronal co-adaptation, dropout layers with a rate of 0.3 were incorporated into fully connected layers. To preserve stable activations and quicken training convergence, batch normalization was used throughout convolutional blocks. A mini-batch gradient descent method with a batch size of 32 was used for the training, guaranteeing effective use of computational resources while preserving gradient stability. The model weights were updated based on the backpropagation of the calculated loss, and each epoch comprised several iterations over the whole training dataset. In order to avoid overfitting and save computation time, early stopping was used with a patience value of 10 epochs to automatically stop training once the validation loss stopped improving.

Monitoring the model’s learning behavior during training was made possible in large part by the validation phase. The model’s ability to generalize beyond the training data was assessed by evaluating its performance on the validation set following each epoch. The model’s discriminative power and dependability were revealed by the recording of validation metrics. In order to identify underfitting or overfitting tendencies and inform choices about regularization and learning rate adjustments, the validation curves which plot training and validation loss over epochs were also examined. In order to preserve the most efficient version of the model for additional testing and explainability integration, model checkpoints were set up to automatically save the network weights corresponding to the best validation accuracy. The CSRNet-based framework was successfully trained to identify minute steganographic distortions while retaining high model stability and interpretability by combining sophisticated optimization techniques, balanced data partitioning, and stringent validation.

* 1. DEEPLIFT XAI IMPLEMENTATION

DeepLIFT was incorporated as a XAI technique to improve the steganalysis framework’s interpretability and transparency. Because of its effectiveness and capacity to link a deep neural network’s output prediction to its input features by comparing each neuron’s activation to a reference (or baseline) activation, DeepLIFT was particularly selected. DeepLIFT functioned as a crucial interpretability layer in this project, explaining why the CSRNet-based model distinguished between cover and stego images. The CSRNet model was fully trained and validated before DeepLIFT was integrated. Because it had already discovered significant spatial and residual patterns linked to steganographic traces, the trained CSRNet served as the baseline network. The DeepLIFT algorithm broke down the final output score into contribution scores for each pixel in the input image by backpropagating the difference between each neuron’s activation and its reference state. This made it easier to determine which areas of the picture had the biggest impact on the model's choice, whether it was to predict it as cover or stego.

Establishing a suitable reference baseline was an essential part of DeepLIFT implementation. An average neutral image obtained from the mean of training images was utilized as the reference input because steganalysis deals with minute noise-like variations in pixel intensities. DeepLIFT was able to successfully capture and contrast variations in pixel activations between cover and stego regions thanks to this decision, which offered a stable baseline. DeepLIFT calculated difference-from-reference scores at each neuron level using this baseline, making sure that each feature's influence on the final output was measured in relation to an unaltered representation. The trained model was run through sample images from both classes in order to perform attribution analysis. Attribution maps, which show the relative significance of pixels or regions that contribute favorably or unfavorably to the classification, were generated by DeepLIFT. The model’s overall decision pattern across samples was summarized by further aggregating these maps using mean and max attribution statistics. While the maximum attributions showed the most important pixel regions that drove the model’s confidence in identifying steganographic embeddings, the mean attributions showed the overall influence of each region.

Starting with the last convolutional block of CSRNet, the implementation extracted attributions layer by layer using the DeepExplain library, which is compatible with TensorFlow and Keras backends. This made it possible to examine spatial and residual features that the network prioritized in greater detail. In order to improve the attribution maps’ interpretability for human analysis, gradient normalization and noise-smoothing techniques were also used to lessen visual artifacts. DeepLIFT transformed the deep model from a black-box classifier into a transparent analytical tool by providing an interpretable mapping between the input image and the model’s output. It made it possible for researchers to determine whether the CSRNet was concentrating on irrelevant textures or significant image regions. Additionally, it made it possible to cross-verify model learning behavior with human domain intuition, guaranteeing the accuracy of the CSRNet’s detection system. DeepLIFT essentially acted as a link between human interpretability and model prediction. It provided a strong framework for comprehending the steganalysis model’s reasoning process by breaking down prediction scores into input-level contributions. In security-sensitive applications, where transparency and confidence in AI-driven decisions are just as vital as accuracy itself, this interpretability is particularly crucial.

**Chapter 4**

**EXPERIMENTAL RESULTS AND DISCUSSION**

* 1. PERFORMANCE OF CSRNET MODEL

To ensure a balanced representation of both classes in each subset, the dataset was split into training, validation, and testing sets using an 80:10:10 split ratio. The Adam optimizer was used for training, with an adaptive learning rate strategy, a categorical cross-entropy loss function, and early stopping to avoid overfitting. Standard classification metrics were used to evaluate the model’s performance. These metrics collectively evaluate the framework’s discriminative power and dependability in differentiating stego images from their cover counterparts. The CSRNet produced outstanding results, demonstrating the model’s strong feature extraction and discriminative ability. The following is a summary of the test performance metrics (Table 1).

1. Result Metrics

|  |  |
| --- | --- |
| Precision | 0.8635 |
| Accuracy | 0.8821 |
| Recall | 0.8713 |
| F1-Score | 0.8616 |

These findings show that the suggested model attained an overall accuracy of 88.21%, demonstrating its potent capacity to generalize and precisely identify hidden information across test samples that have not yet been seen. The model strikes a good balance between detecting real stego instances and preventing false positives, as evidenced by the close alignment of Precision and Recall values. Hyperparameter tuning was done by experimenting with parameters like learning rate, batch size, dropout rate, and number of convolutional filters in order to further improve model performance. The model’s convergence rate was increased and training was stabilized thanks to the optimization process. The main hyperparameters investigated during tuning and their ideal values are compiled below (Table 2).

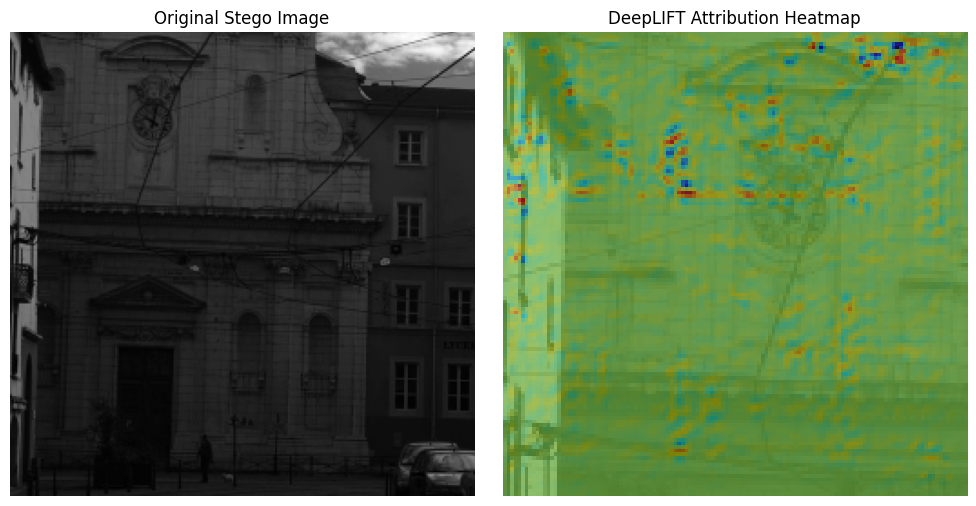
1. Hyperparameter Tuning Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Experiment** | **Train-Test- Validation Split** | **Learning Rate** | **Batch Size** | **Validation Accuracy** |
| 1 | 60-20-20 | 0.005 | 32 | 0.55 |
| 2 | 70-20-10 | 0.002 | 8 | 0.67 |
| 3 | 70-20-10 | 0.001 | 16 | 0.74 |
| 4 | 80-10-10 | 0.005 | 16 | 0.79 |
| 5 | 80-10-10 | 0.001 | 8 | 0.88 |

The model’s stability was confirmed by the training and validation accuracy curves, which showed consistent improvement with little overfitting. Effective learning and generalization capacity were further demonstrated by the loss curve’s steady decline over epochs. Additionally, the AFRU units used residual and attention-based mechanisms to focus attention on the most informative areas of the image, while the use of SRM filter banks improved the model’s capacity to capture high-frequency noise and texture details introduced by steganographic embedding. Additionally, by identifying areas most likely to contain hidden modifications, the CSM greatly improved performance. By incorporating this selective mechanism, CSRNet was able to focus on minute pixel-level anomalies while effectively ignoring uninformative regions. CSRNet is a reliable and comprehensible solution for contemporary steganalysis problems because of its consistent performance across several metrics and its interpretability via DeepLIFT XAI. These findings confirm that the suggested architecture can reliably and accurately distinguish between cover and stego images, even in the presence of subtle embedding patterns that frequently elude conventional detection techniques.

* 1. PERFORMANCE OF DEEPLIFT XAI

The trained CSRNet model was subjected to DeepLIFT to guarantee transparency and interpretability in the steganalysis process. DeepLIFT is a potent XAI method that determines how each input feature (pixel) contributes to the final decision made by the model. It provides important insights into the network’s internal reasoning process in the context of steganalysis by making it possible to visualize and comprehend which image regions influenced the classification as cover or stego. In order to guarantee that the attribution scores were computed in relation to the original state of the image, the CSRNet model was used as the baseline reference for DeepLIFT computations. DeepLIFT produced attribution maps (heatmaps) that graphically depict the degree and polarity of each pixel’s influence on the classification decision for every test image. While regions with negative attributions indicate areas that supported classification as cover, regions with high positive attributions correspond to areas that significantly contributed to identifying an image as stego (Fig. 7).



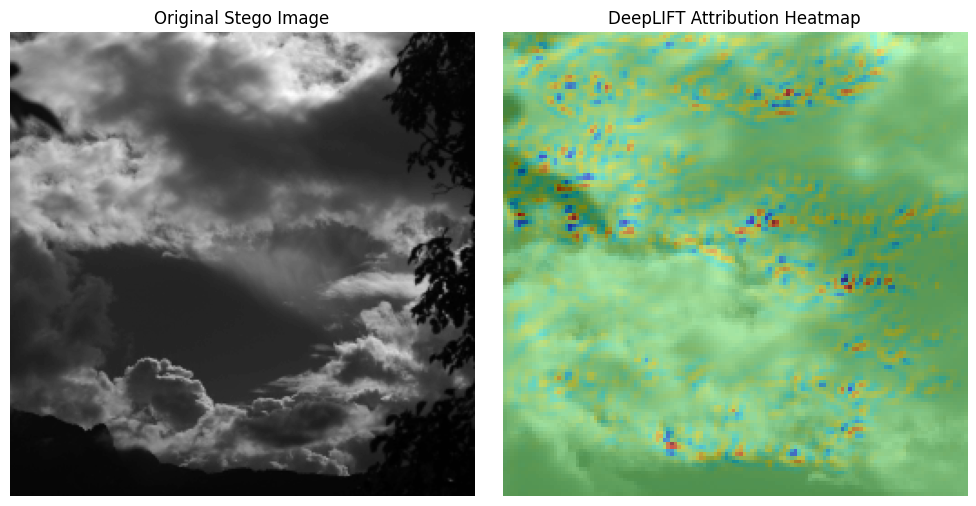


Fig.7 (DeepLift Attribution Heatmap for Stego Images)

To quantify feature importance across the test samples, two main statistical measures were extracted, mean attribution and top influential attributions. While top attributions indicated the particular areas that had the biggest effects on the model’s output, mean attribution values gave a general idea of how average pixel intensity variations affected the prediction. High-frequency noise, subtle edge distortions, and fine-grained irregularities, that is patterns commonly introduced during steganographic embedding were consistently associated with the top attributions in correctly classified stego images. On the other hand, attribution patterns in cover images were more consistent and spatially smooth, suggesting that there were no such disruptions (Fig. 8).

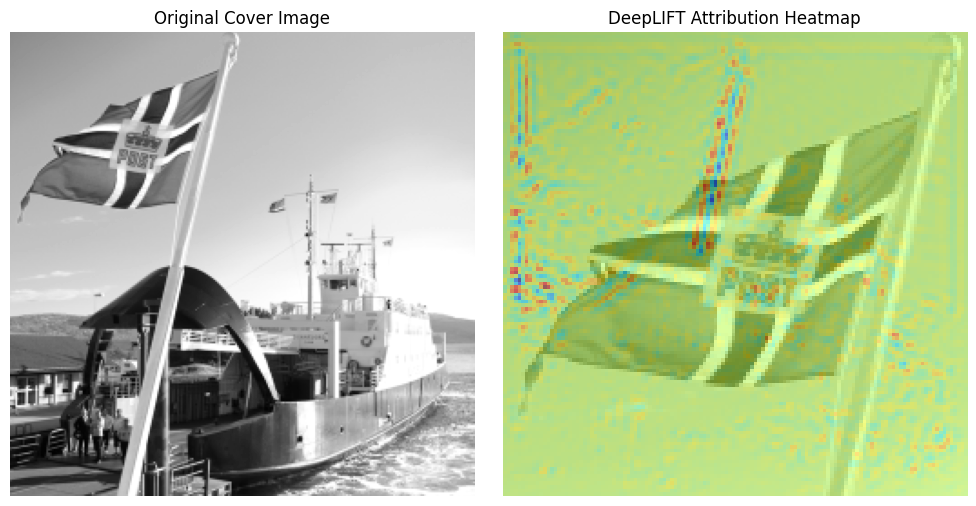




Fig.8 (DeepLift Attribution Heatmap for Cover Images)

The heatmaps produced by DeepLIFT clearly outperformed traditional feature visualization techniques in terms of interpretability. They showed that CSRNet was trained to concentrate on localized anomalies, especially in the vicinity of object boundaries and textured regions where embedding algorithms frequently conceal information. By focusing the network’s attention on significant areas rather than random noise, the adaptive behavior of the AFRU and the selective attention directed by the CSM further strengthened this interpretability. Furthermore, the DeepLIFT attributions validated the reliability of the model by confirming that CSRNet’s decisions were based on visually explainable evidence rather than being arbitrary. The idea that DeepLIFT successfully reveals embedding patterns that are otherwise undetectable to the human eye is supported, for example, by comparing a cover and its corresponding stego image heatmaps, which clearly showed increased activity around the modified pixels. Overall, the DeepLIFT results highlight the explainability and diagnostic power of the framework. The suggested CSRNet+DeepLIFT combination achieves high classification accuracy (88.21%) and offers clear explanations for each prediction by linking pixel-level feature importance with model predictions. For practical forensic applications, this interpretability is essential because it allows human analysts to confirm and trust the results of AI-driven steganalysis.

**Chapter 5**

**CONCLUSION AND FUTURE WORK**

* 1. CONCLUSION

Digital forensics and cybersecurity have advanced significantly with the proposed Explainable Steganalysis Framework using CSRNet with Adaptive SRM Filtering and Integration of XAI. This study effectively bridges the gap between model interpretability and high-performance steganalysis, two areas that are frequently addressed separately in earlier research. By combining DeepLIFT XAI with CSRNet, the framework improves transparency and confidence in AI-driven analysis by achieving strong detection accuracy and offering concise visual explanations for its predictions. By dynamically modifying attention based on temperature scaling, the CSM implementation proved essential in identifying possible stego regions. The model’s ability to selectively focus on areas that might contain hidden information was made possible by this adaptive mechanism, which increased detection accuracy. The model’s capacity to capture minute residual patterns and spatial artifacts introduced during steganographic embedding was improved by the use of five SRM filter banks. Additionally, by utilizing residual connections and attention-guided filtering, the AFRU improved feature extraction, guaranteeing efficient gradient propagation and minimizing information loss during training.

The CSRNet’s robustness across a variety of image samples was validated through an experimental accuracy of 88.21% with balanced precision and recall values. By providing pixel-level attribution maps that graphically explain why a specific image was categorized as cover or stego, the integration of DeepLIFT XAI added a crucial layer of interpretability. Because forensic analysts need both performance and interpretive evidence to make decisions, these explainability insights increase the system’s dependability. This work’s overall contribution is the creation of an end-to-end explainable steganalysis model that integrates interpretability mechanisms, adaptive filtering, and deep feature learning. By offering useful insights into the network’s decision-making process, this method goes beyond conventional black-box steganalysis techniques. The framework’s practical applications go beyond scholarly experimentation. Real-world forensic tools that can analyze image-based data exchanges across digital communication platforms can be built upon it. Its interpretability can help law enforcement, cybersecurity analysts, and digital investigators identify and visualize steganographic threats in online media. In addition to improving steganalysis’s technical capabilities, the project helps meet the increasing demand for transparent and moral AI systems in cybersecurity applications.

* 1. FUTURE WORK

Although the suggested explainable steganalysis framework based on CSRNet shows strong performance and interpretability, there is still a lot of room for improvement and wider use. In order to improve the model’s adaptability, scalability, and applicability across various digital ecosystems, future research directions can investigate improvements at the architectural, dataset-level, and real-world deployment levels. Applying the framework to real-time and cross-platform steganalysis, especially in social media and instant messaging settings, is a crucial extension of this work. Rapid image sharing on platforms like Facebook Messenger, Telegram, and WhatsApp can be used by bad actors to steganographically embed malware, phishing payloads, or private information in images. This hidden threat is unknown to the majority of users and even to current moderation tools. In order to prevent potential data loss, privacy violations, or device compromises, the suggested CSRNet model could be used as a background forensic monitoring tool to identify and flag such embedded content before it reaches end users. The model could be modified to create lightweight, cloud-based APIs that can instantly analyze uploaded images in order to accomplish this integration.

Future research can concentrate on improving model generalization across various datasets, image formats, and embedding algorithms. Expanding to color datasets like ALASKA2 or ImageNet-based stego corpora would enable the model to handle more intricate spatial and chromatic features. The current study used the BOSSBase dataset, which mainly consists of grayscale images. Furthermore, investigating transformer-based attention mechanisms or self-supervised feature learning may enhance the model’s capacity to detect minute embedding traces without significantly depending on labeled data. Combining CSRNet with multi-modal explainable AI methods, like Layer-wise Relevance Propagation (LRP) or Integrated Gradients, to compare and validate feature importance from various interpretability perspectives is another promising avenue. This would improve the framework’s dependability and auditability for use in legal and forensic investigations.

Furthermore, by employing methods like model pruning, quantization, and knowledge distillation, the model could be optimized for hardware-efficient deployment, enabling integration into mobile applications or edge devices. This could add another level of digital security by enabling on-device detection of steganographic content prior to files being downloaded or shared. In conclusion, the research’s future scope is multifaceted, spanning from theoretical developments in explainable AI and steganalysis to real-world applications in digital security ecosystems. By enabling proactive detection of hidden malicious data and guaranteeing safer, more transparent communication networks, extending this work toward social media and messaging platform integration can have a significant real-world impact.

* 1. CHALLENGES FACED

Several difficulties arose at different phases of the implementation of the suggested explainable CSRNet-based steganalysis framework. The fact that steganographic modifications are intrinsically subtle, introducing only minute residual changes that are almost undetectable to both the human eye and conventional statistical methods, was one of the main challenges. Another difficulty was maintaining these high-frequency features during deep learning processing because standard convolutional operations frequently suppress patterns that resemble noise, requiring the careful integration of SRM filter banks to improve residual extraction. The need for exact temperature tuning to strike a balance between excessively soft and excessively discrete attention maps for efficient patch localization made the CSM more complicated.

It took several iterations to ensure stable gradient flow and feature refinement when designing the AFRU to selectively gate pertinent features while incorporating depth-wise separable convolutions, normalization, dropout, and residual connections. Even with the balanced dataset, it was difficult to avoid overfitting because pixel-level feature sensitivity and repeated augmentations could lead to performance instability. To guarantee that attribution maps were significant and correlated with real stego-affected areas, the integration of DeepLIFT XAI also required careful baseline selection. Finally, the combination of SRM filtering, AFRU layers, and attention mechanisms increased computational load, resulting in longer training times and requiring resource allocation and batch size optimization to maintain efficiency without sacrificing performance.

* 1. LIMITATIONS

Although the suggested Explainable CSRNet-based steganalysis framework achieves strong performance and interpretability, there are some limitations that offer room for further study and improvement. These restrictions include explainability depth, computational complexity, model generalization, and dataset limitations. First, 10,000 cover and 10,000 stego grayscale images from the BOSSBase dataset were used to train and validate the model. Although the steganalysis community uses this dataset as a benchmark, it only covers a small portion of grayscale images and natural photographic content. As a result, when the model is tested on color images, social media images, or datasets with varying compression levels and noise characteristics, its performance may deteriorate. This draws attention to a possible restriction in the CSRNet framework’s capacity to generalize to real-world and cross-domain situations.

Second, applying several SRM filters and training CSRNet come at a comparatively high computational cost. Because the SRM filter bank uses convolution with high-dimensional kernels, preprocessing time and GPU memory usage are greatly increased. This makes on-device or real-time deployment difficult, particularly in settings like mobile or embedded systems that have constrained processing power. Furthermore, because CSRNet has a lot of parameters, stable convergence requires careful hyperparameter tuning, which can be costly and time-consuming. The integrated DeepLIFT XAI module’s interpretability depth is another drawback. DeepLIFT does not always offer high-level semantic explanations for why particular patterns correspond to steganographic artifacts, even though it is successful in identifying pixel-level attributions that affect model predictions. To completely comprehend the embedding traces, expert analysis is necessary as the visualization of attributions is still subject to subjective interpretation. Furthermore, DeepLIFT relies on the baseline model (CSRNet) being well-calibrated; bias or overfitting in CSRNet may spread to the explanation results, impacting interpretability reliability.

Additionally, the data augmentation techniques used, such as random resizing, flipping, brightness, and contrast adjustments may unintentionally change the delicate pixel correlations that are essential for identifying subtle steganographic signals, even though their main goal is to increase model robustness. As a result, the model may occasionally learn augmentation-induced noise rather than actual embedding patterns. Therefore, it is still difficult to strike the ideal balance between augmentation diversity and feature preservation. Lastly, even though the model’s accuracy is 88%, there is still room for improvement when it comes to identifying sophisticated adaptive steganographic methods like HILL, WOW, and S-UNIWARD, which strategically distribute payloads to reduce statistical disruptions. The model’s limited sensitivity to changes in payload size and embedding algorithms limits its applicability to more complex or encrypted steganography techniques. In conclusion, even though the suggested CSRNet with DeepLIFT XAI framework shows encouraging accuracy and interpretability, problems with dataset diversity, computational scalability, real-time adaptability, and explainability robustness can improve its performance even more. These drawbacks offer important avenues for further study to improve the effectiveness, generalizability, and transparency of steganalysis models in practical cybersecurity applications.

**APPENDICES**

1. **- Dataset Information**

The BOSSBase Image Steganalysis Dataset, one of the most popular benchmark datasets for assessing steganographic detection models, was used in the experimental evaluation.

* Twenty thousand grayscale photos in total
* Classifications:
  + Unaltered cover photos: 10,000
  + 10,000 Stego photos (embedded)
* Format: Portable Gray Map, or PGM
* 512 x 512 pixels in resolution
* Color Mode: single-channel grayscale
* Data Division:
  + 80% of the training set
  + 10% is the validation set.
  + 10% is the testing set.

The dataset ensures diversity in textures, lighting, and spatial patterns by representing a wide range of natural scenes. The model is better able to generalize to real-world images with various embedding schemes thanks to this variability.

1. **- Hyperparameter Tuning**

Several hyperparameters were adjusted during testing to get the best model performance. The following is a summary of the final setup used for CSRNet training:

1. Hyperparameter Configuration

|  |  |  |
| --- | --- | --- |
| **Parameter** | **Description** | **Value** |
| Learning Rate | Controls step size during optimization | 0.0005 |
| Optimizer | Algorithm used for gradient updates | Adam |
| Batch Size | Number of samples per training batch | 32 |
| Dropout Rate | Regularization rate to prevent overfitting | 0.4 |
| Epochs | Total number of training iterations | 80 |
| Activation Function | Non-linear transformation applied to neurons | ReLU |
| Loss Function | Error function used for classification | Binary Cross-entropy |
| Metrics | Evaluation parameters | Accuracy, Precison, Recall, F1-score |
| Validation Split | Fraction of data reserved for validation | 0.1 |

1. **- Model Architecture Details**

The Three essential elements are integrated into the CSRNet architecture:

* Module for Cover Selection (CSM):
  + Finds possible steganographic embedding regions.
  + Controls the "softness" of attention using a temperature parameter.
  + Uses a sigmoid activation and three convolutional layers to create selection maps.
* SRM Filter Banks:
  + Comprises five spatial filters: Sobel-X, Sobel-Y, Laplacian, Laplacian-8, and High-Pass.
  + Improves pixel-level irregularities, residual noise, and edge details.
* Adaptive Filtered Residual Unit (AFRU):
  + Uses the CSM output to process residual feature maps and implement attention-guided modulation.
  + Makes use of dropout regularization, ReLU activations, and depth-wise separable convolutions.
  + Incorporates skip connections for better training convergence and stable gradient flow.

1. **- Experiment Environment**

A local high-performance computing setup with the following configuration was used for all experiments. Effective training, visualization, and interpretability analysis were made possible by this configuration. SRM filtering, residual extraction, and DeepLIFT attribution mapping computation times were greatly shortened by GPU acceleration.

1. Experiment Environment Details

|  |  |
| --- | --- |
| **Component** | **Specification** |
| Operating System | Windows 11 |
| Programming Language | Python 3.10 |
| Deep Learning Framework | Pytorch |
| Hardware Accelerator | NVIDIA RTX 3060 GPU (12 GB VRAM) |
| CPU | Intel Core i7 12th Gen |

1. **– Evaluation Details**

The following evaluation metrics were used to gauge the model’s performance:

* The overall percentage of correctly classified images is known as accuracy (Acc).
* The ratio of accurate stego predictions to all stego predictions is known as precision (P).
* Recall (R): The proportion of accurately recognized stego images to all stego images.
* F1-Score: A balanced performance metric derived from the harmonic mean of precision and recall.

The confusion matrix obtained from test predictions was used to calculate these metrics. Three essential elements are integrated into the CSRNet architecture

1. **– DeepLIFT Configuraion**

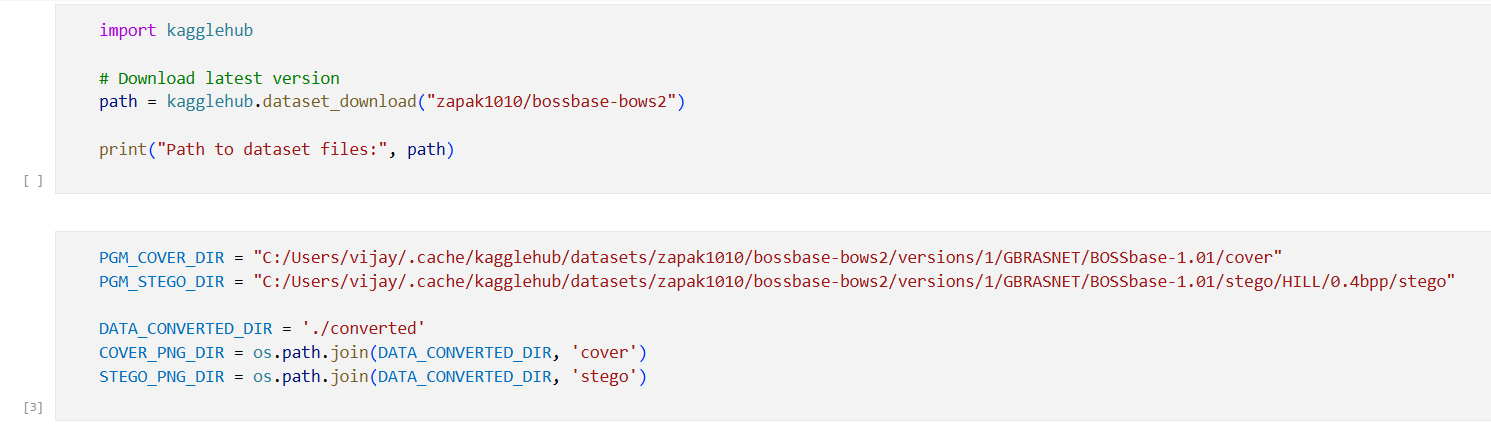
For explainability analysis, DeepLIFT was incorporated after training.

* Baseline Model: CSRNet with training.
* Top Influence, Max Attribution, and Mean Attribution are the methods used for attribution.
* Heatmaps displaying the positive and negative contributions for each pixel in the output visualization.
* Interpretation:
  + Stego embedding is strongly indicated by high positive attribution regions.
  + Cover classification is supported by regions with negative attribution.
  + Pixels with neutral attribution have little bearing on prediction.

The produced heatmaps validated CSRNet’s capacity to identify minute steganographic changes by offering qualitative insights into pixel-level significance.

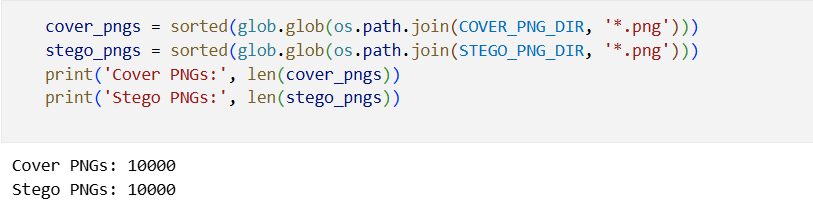
**CODE SNIPPETS**

1. **Dataset Importation**

****

1. **PGM to PNG Conversion**

****

****

1. **Data Labeling and Pre-processing**

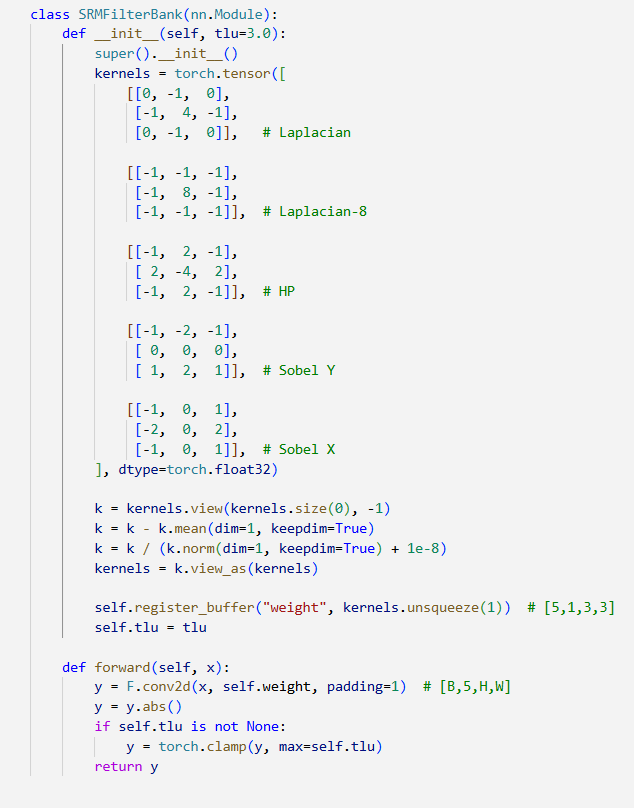
****

****

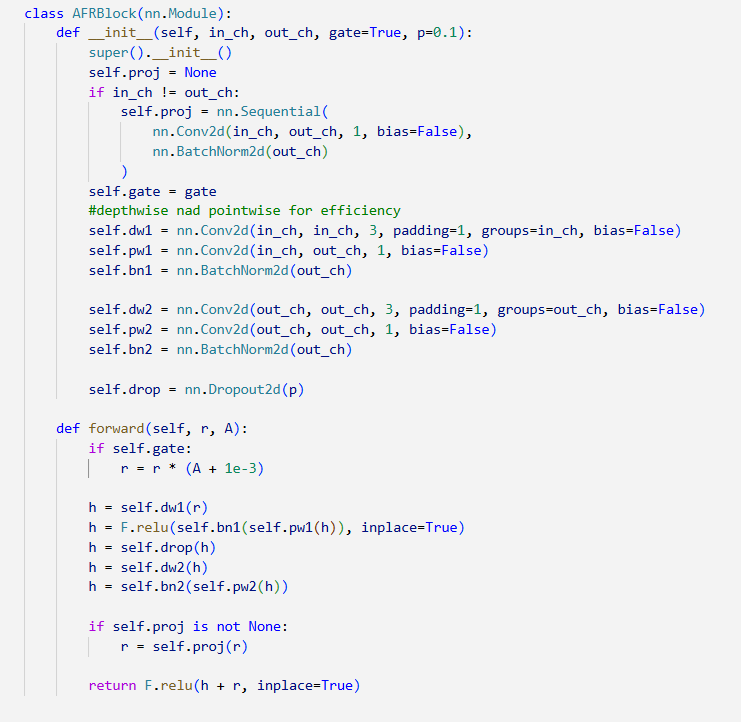
1. **Cover Selection Module (CSM)**

****

1. **SRM Filterbanks**

****

1. **Adaptive Filtered Residual Block (AFRU)**

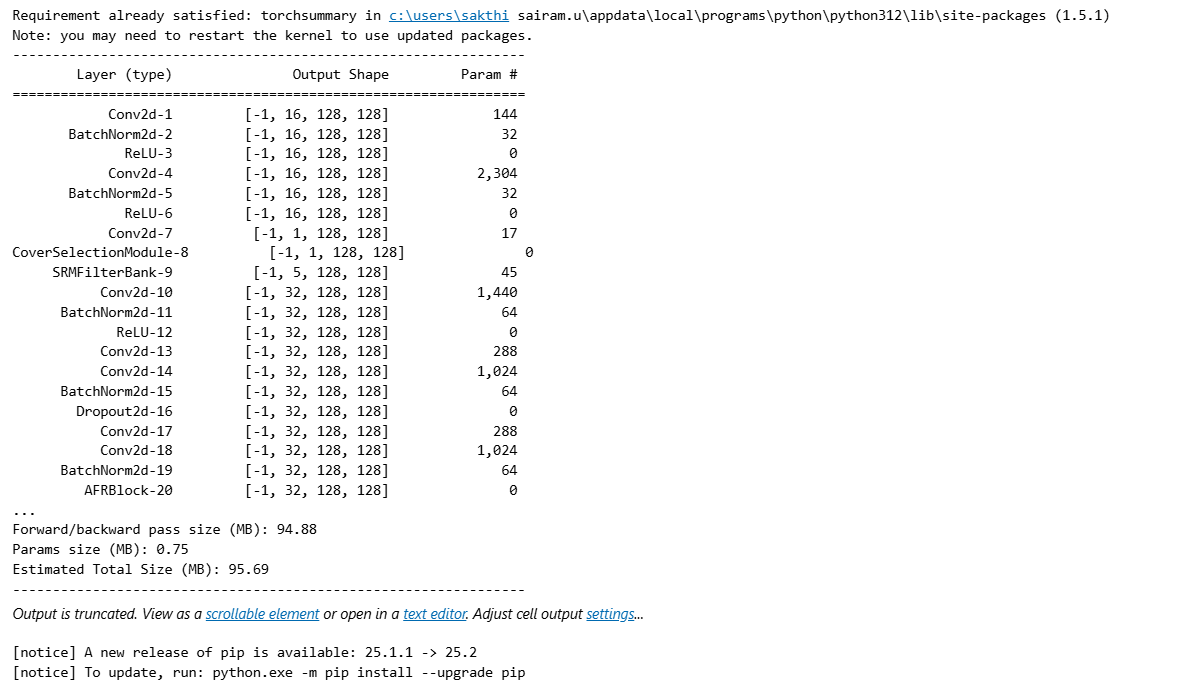
****

1. **CSRNet**

****

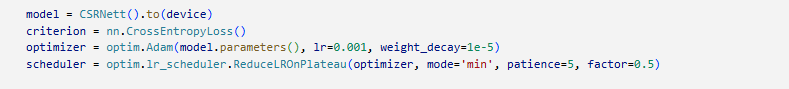
1. **CSRNet Summary**

****

****

1. **Dataset Transformation and Splitting**

****

****

1. **Model Training**

****

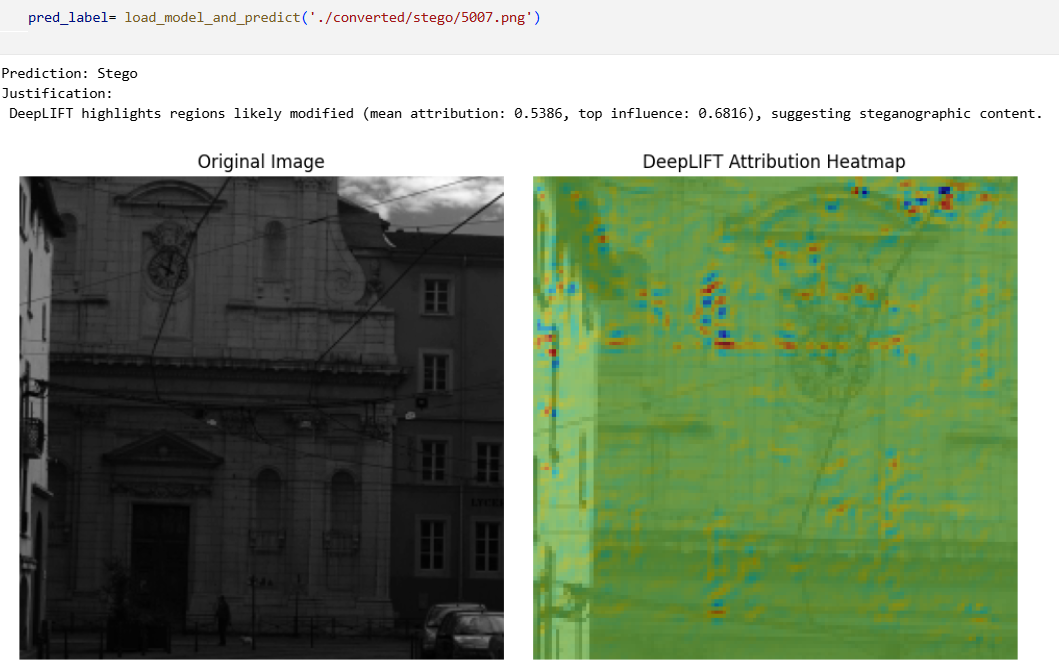
1. **Model Evaluation**

****

1. **Model Testing**

****

****

****

**REFERENCES**

1. Ma, Y., Yang, Z., Li, T., Xu, L., & Qiao, Y. (2023). Image steganalysis method based on cover selection and adaptive filtered residual network. Computers & Graphics, 115, 43-54.
2. Kheddar, H., Hemis, M., Himeur, Y., Megias, D., & Amira, A. (2024). Deep learning for steganalysis of diverse data types: A review of methods, taxonomy, challenges and future directions. Neurocomputing, 581, 127528.
3. Gupta, A., Chhikara, R., & Sharma, P. (2024). CIRNet: An improved lightweight convolution neural network architecture with inverted residuals for universal steganalysis. Arabian Journal for Science and Engineering, 49(9), 12219-12233.
4. Gao, X., Yi, J., Liu, L., & Tan, L. (2025). A generic image steganography recognition scheme with big data matching and an improved ResNet50 deep learning network. Electronics, 14(8), 1610.
5. Zhu, Y., Wang, X., Chen, H. S., Salloum, R., & Kuo, C. C. J. (2023). Green steganalyzer: A green learning approach to image steganalysis. APSIPA Transactions on Signal and Information Processing, 12(1).
6. Shehab, D. A., & Alhaddad, M. J. (2022). Comprehensive survey of multimedia steganalysis: Techniques, evaluations, and trends in future research. Symmetry, 14(1), 117.
7. Kuznetsov, O., Frontoni, E., Chernov, K., Kuznetsova, K., Shevchuk, R., & Karpinski, M. (2024). Enhancing Steganography Detection with AI: Fine-Tuning a Deep Residual Network for Spread Spectrum Image Steganography. Sensors, 24(23), 7815.
8. Huang, S., Zhang, M., Ke, Y., Bi, X., & Kong, Y. (2022). Image steganalysis based on attention augmented convolution. Multimedia Tools and Applications, 81(14), 19471-19490.
9. Mhawi, M. Y., Abdullah, H. N., & Sikora, A. (2024). Gans for image security applications: a literature review. Iraqi Journal of Information and Communication Technology, 7(2), 89-102.
10. De La Croix, N. J., Ahmad, T., & Han, F. (2024). Comprehensive survey on image steganalysis using deep learning. array, 22, 100353.
11. Bohang, L., Li, N., Yang, J., Alfarraj, O., Albelhai, F., Tolba, A., ... & Yee, P. L. (2025). Image steganalysis using active learning and hyperparameter optimization. Scientific Reports, 15(1), 7340.
12. Eid, W. M., Alotaibi, S. S., Alqahtani, H. M., & Saleh, S. Q. (2022). Digital image steganalysis: current methodologies and future challenges. Ieee Access, 10, 92321-92336.
13. Farooq, N., & Selwal, A. (2023). Image steganalysis using deep learning: a systematic review and open research challenges. Journal of Ambient Intelligence and Humanized Computing, 14(6), 7761-7793.
14. Shrikumar, A., Greenside, P., & Kundaje, A. (2017, July). Learning important features through propagating activation differences. In International conference on machine learning (pp. 3145-3153). PMlR.
15. Kuchumova, E., Martínez-Monterrubio, S. M., & Recio-Garcia, J. A. (2024). STEG-XAI: explainable steganalysis in images using neural networks. Multimedia Tools and Applications, 83(17), 50601-50618.
16. Chahar, N. K., Dhaka, A., Nandal, A., & Kumar, V. (2025, June). An Explainable Deep Learning Framework for Usable and Secure Image Steganography. In 2025 International Conference on Electronics, AI and Computing (EAIC) (pp. 1-6). IEEE.
17. Farooq, N., & Mir, R. N. (2024). Image Steganalysis using Deep Convolution Neural Networks: A Literature Survey. International Journal of Sensors, Wireless Communications and Control, 14(4), 247-264.
18. Chang, C. C., Wang, X., Chen, S., Echizen, I., Sanchez, V., & Li, C. T. (2023). Deep learning for predictive analytics in reversible steganography. IEEE Access, 11, 3494-3510.
19. Kraetzer, C., & Hildebrandt, M. (2024). Explainability and Interpretability for Media Forensic Methods: Illustrated on the Example of the Steganalysis Tool Stegdetect. Proceedings Copyright, 585, 592.
20. Hu, M., & Wang, H. (2025). Mutual Information-Optimized Steganalysis for Generative Steganography. IEEE Transactions on Information Forensics and Security.
21. Chang, C. C., Wang, X., Chen, S., Echizen, I., Sanchez, V., & Li, C. T. (2023). Deep learning for predictive analytics in reversible steganography. IEEE Access, 11, 3494-3510.