Adversarial-robust steganalysis system leveraging adversarial training and EfficientNet

Thakwan Akram Jawad[1,2], Jamshid Bagherzadeh Mohasefi[1], Mohammed Salah Reda Abdelghany[3]
[1]Department of Computer Engineering, Faculty of Electrical and Computer Engineering, Urmia University, Urmia, Iran
[2]Department of System and Control Engineering, College of Electronic Engineering, Ninevah University, Mosul, Iraq
[3]Department of computer science, Faculty of computers and artificial intelligence, Beni-Suef University, Beni Suef, Egypt

Article Info

Article history:

Received Aug 24, 2024
Revised Dec 27, 2024
Accepted Jan 22, 2025

Keywords:

Adversarial attacks
Deep learning
EfficientNet
Steganalysis
Steganography

ABSTRACT

Steganalysis aims to detect hidden messages within digital media, presenting a significant challenge in the field of information security. This paper introduces an adversarial-robust steganalysis system leveraging adversarial training and the powerful feature extraction capabilities of EfficientNet. We utilize EfficientNet to extract robust features from images, which are subsequently classified by a dense neural network to distinguish between steganographic and non-steganographic content. To enhance the system's resilience against adversarial attacks, we implement a custom adversarial training loop that generates adversarial examples using the fast gradient sign method (FGSM) and integrates these examples into the training process. Our results demonstrate that the proposed system not only achieves high accuracy in detecting steganographic content but also maintains robustness against adversarial perturbations. This dual approach of leveraging state-of-the-art deep learning architectures and adversarial training provides a significant advancement in the field of steganalysis, ensuring more reliable detection of hidden messages in digital images.

Corresponding Author:

Jamshid Bagherzadeh Mohasefi
Department of Computer Engineering, Faculty of Electrical and Computer Engineering, Urmia University
11 Km Serow Road, Urmia, West Azerbaijan, Iran
Email: j.bagherzadeh@urmia.ac.ir

1. INTRODUCTION

In today's digital age, ensuring the security of information transmission has become a critical concern. Steganography, the practice of hiding secret messages within digital media, poses a unique challenge as it can be used for both legitimate and malicious purposes. Steganalysis, the art of detecting these hidden messages, is therefore of paramount importance [1]-[3]. Traditional steganalysis methods often struggle with robustness, especially when faced with adversarial attacks designed to deceive detection systems [4], [5].

The need to enhance model robustness against adversarial attacks has become a critical area of focus in the field of machine learning. Adversarial attacks pose significant threats by subtly altering input data, which can deceive models into making erroneous predictions, potentially leading to severe consequences in sensitive domains like finance, healthcare, and autonomous systems [6]. These attacks exploit the vulnerabilities in a model's decision boundary, revealing that models might perform well in controlled environments but are fragile when faced with adversarial inputs. Strengthening the robustness of models is essential to ensure their reliability and accuracy in real-world applications, where adversarial conditions might be encountered [7]. Approaches such as adversarial training, which involves training models with adversarial examples, and the use of robust optimization techniques, have shown promise in mitigating these vulnerabilities [8]. Additionally, implementing defensive strategies like gradient masking and input transformation can further enhance the model’s ability to withstand adversarial perturbations [9]. By focusing on developing robust machine learning models, the field can better protect these systems from adversarial threats, ensuring their safety and effectiveness across various applications.

In this paper, we propose an advanced steganalysis system that leverages the feature extraction capabilities of EfficientNet [10] and incorporates adversarial training to enhance robustness against adversarial attacks. Our approach involves training a dense neural network on features extracted by EfficientNet, combined with a custom adversarial training loop using the fast gradient sign method (FGSM) [11]. Goodfellow *et al.* [11] introduced the FGSM to create adversarial examples. This technique leverages the derivative of the network’s loss function concerning the input feature vector. For a given input image, FGSM perturbs each feature by adjusting it in the direction of the gradient. Consequently, this modification alters the classification outcome of the input image. For a neural network with a cross-entropy loss function, where the input image and target class are specified, the adversarial example is generated by adding a small perturbation to the original image. This perturbation is controlled by a parameter that determines its size, ensuring the changes are subtle yet effective in misleading the network [11].

DeepFool is an untargeted attack method that generates adversarial examples through iterative perturbations of an image. This approach aims to find the nearest decision boundary by slightly modifying the image in each iteration until the altered image changes the network’s classification [12]. The Carlini and Wagner (C&W) method, named after its creators, is a powerful attack that can be either targeted or untargeted. It utilizes three metrics to measure distortion: L0, L2, and L∞ norms, with the untargeted L2 norm version being noted for its superior performance. This method generates adversarial examples by solving an optimization problem to find the smallest perturbation that misleads the network. The optimization balances minimizing the perturbation and altering the network’s classification, controlled by a hyperparameter. This attack seeks the minimal L2 norm perturbation to change the image’s classification while ensuring the perturbed image remains within valid pixel values. The method also incorporates a confidence parameter to generate adversarial examples with high classification confidence, which enhances their transferability to other models. The C&W method is recognized for its robustness and difficulty to defend against [13].

Qin *et al.* [14] leverages vector-wise embedding with color pixel vectors (CPV) for color image steganography, improving upon traditional element-wise embedding. In particular, this method uses the iterative fast gradient sign method (I-FGSM) to generate adversarial perturbations that effectively deceive steganalytic convolutional neural networks (CNNs). By incrementally adjusting adversarial costs within non-overlapping sub-images, this approach reduces detectability, maintaining a high success rate against CNN-based steganalyzers. Robustness-based defense focuses on correctly classifying adversarial examples, and various methods have been developed to achieve this. One common approach is adversarial training, which involves augmenting the training set with a mix of normal and adversarial examples to improve the network’s resilience [11].

Another strategy is preprocessing input images to remove adversarial perturbations through techniques such as principal component analysis (PCA). Bhagoji *et al.* [15] investigate the enhancement of machine learning system robustness through various data transformations. Their study reveals that certain transformations, such as input preprocessing, feature manipulation, and PCA, can significantly mitigate the impact of adversarial attacks, thereby improving the resilience of neural networks to adversarial perturbations [15], [16].

Another method is JPEG compression, adding noise, cropping, and rotating. Das *et al.* [17] propose using JPEG compression as a preprocessing step to defend against adversarial attacks. By removing high-frequency signal components through selective blurring, JPEG compression can mitigate the impact of perturbations added to images. Their approach shows that this technique can significantly reduce the efficacy of several attack methods, including the FSGM and DeepFool, without requiring prior knowledge of the model under attack.

Papernot *et al.* [18] introduce defensive distillation as a method to enhance the robustness of deep neural networks (DNNs) against adversarial attacks. By leveraging knowledge distillation techniques, this approach reduces the effectiveness of adversarial sample generation, significantly mitigating the success rate of such attacks. The study demonstrates a dramatic reduction in adversarial effectiveness, highlighting the potential of distillation to improve DNN security and resilience. Defensive distillation enhances robustness by hiding the gradient between the pre-softmax layer and softmax outputs usinqing distillation training techniques [16], [18].

Additionally, obfuscated gradients hinder attackers by making it difficult to compute feasible gradients for generating adversarial examples. Athalye *et al.* [19] identify and analyze obfuscated gradients, a form of gradient masking that falsely appears to offer security against adversarial attacks. The study demonstrates that defenses relying on obfuscated gradients can be effectively circumvented, showcasing techniques to bypass such defenses and revealing their vulnerability in a white-box setting. The paper emphasizes that many defenses accepted at International Conference on Learning Representations (ICLR)

2018, which claim robustness, actually depend on obfuscated gradients, making them susceptible to targeted attacks [19].

Meng and Chen [20] proposed MagNet, a framework for defending neural networks against adversarial examples. This framework neither modifies the protected classifier nor requires knowledge of the adversarial example generation process. MagNet employs detector networks to distinguish between normal and adversarial examples by approximating the manifold of normal examples. Additionally, a reformer network is used to move adversarial examples closer to the manifold of normal examples, enabling correct classification. MagNet has demonstrated effectiveness against state-of-the-art attacks in black-box and gray-box scenarios without sacrificing false positive rates.

Taran *et al.* [21] proposed a key-based diversified aggregation (KDA) mechanism to defend DNNs against adversarial attacks in gray-box and black-box scenarios. KDA involves a multi-channel architecture with randomized transformations applied to each channel. The randomization, based on a secret key, prevents gradient backpropagation and hinders the attacker's ability to create effective adversarial examples. By aggregating the outputs from multiple channels, KDA improves the robustness and reliability of the classification process. The authors demonstrated the effectiveness of KDA against various state-of-the-art attacks, including gradient-based and non-gradient-based methods [21].

EfficientNetB0 is part of the EfficientNet family, a series of CNNs designed for high performance and efficiency. The key innovation in EfficientNet is the use of a compound scaling method that uniformly scales all dimensions of depth, width, and resolution using a simple yet effective scaling coefficient. EfficientNetB0, the baseline model, achieves remarkable accuracy and efficiency by optimizing this balance. It utilizes fewer parameters and computational resources while maintaining or even improving performance compared to previous architectures. EfficientNetB0 employs a combination of depth wise convolutions, squeeze-and-excitation optimization, and the MBConv block (a variant of inverted residual blocks). These architectural choices allow it to achieve state-of-the-art performance on several benchmarks while being significantly more computationally efficient [10].

Random forest is an ensemble learning method that combines multiple decision trees to improve classification accuracy and robustness [22]. This model is known for its ability to handle large datasets and perform well with unstructured data, making it a popular choice for various machine learning tasks. However, random forests are susceptible to adversarial examples, which are intentionally perturbed inputs designed to deceive machine learning models. Unlike DNNs, which can learn complex representations and potentially develop some level of resilience to adversarial attacks through specialized training techniques, random forests lack inherent mechanisms to deal with such perturbations effectively. Adversarial examples exploit the linear decision boundaries of each individual tree within the ensemble, leading to misclassifications. Recent studies have shown that adversarial attacks can reduce the accuracy of random forest models significantly more than that of well-trained DNNs, which can be fortified through adversarial training and regularization techniques [18], [23].

Therefore, while random forests provide a simple and interpretable model, their vulnerability to adversarial attacks highlights the need for additional defenses when deploying these models in adversarial settings. In this paper, we contribute the following novelties:

- We implemented three different machine learning and deep learning models on our dataset and evaluated the models under attack.
- We used EfficientNet to extract features from images and trained the models above on these extracted data.
- We achieved high accuracy and precision in our proposed method compared to other base models.

This paper is organized as follows. In section 2, we explain the data preparation and preprocessing operations. Then in section 3, we explain the proposed methods including deep learning and machine learning techniques. Later, we show the results of the implemented methods in section 4. Finally, section 5 presents the discussion and concludes the paper.

2. METHOD

Our proposed method integrates EfficientNet for feature extraction and adversarial training to enhance the robustness of a steganalysis system. EfficientNet is chosen for its ability to capture intricate patterns in image data with fewer parameters, making it both effective and computationally efficient. The model is trained using a dataset augmented with adversarial examples generated through the FGSM, which introduces slight perturbations to the input data. This process helps the model learn to maintain high performance even when facing adversarial attacks. The architecture of our model includes layers such as dropout and batch normalization to improve generalization and stability by reducing overfitting and normalizing feature inputs, respectively. These layers are critical in ensuring that the model remains robust under various conditions, particularly in adversarial settings. Compared to traditional machine learning approaches like random forest, our model offers superior accuracy and resilience, demonstrating its capability to not only accurately classify images but also withstand adversarial perturbations, which is essential for real-world applications where security and reliability are paramount.

EfficientNet is a high-performance model architecture optimized for computational efficiency. It scales depth, width, and resolution in a compound manner, enabling smaller models to achieve higher accuracy than many conventional architectures [10]. In adversarial training, a model is iteratively exposed to adversarially perturbed inputs during training to improve its robustness. This method uses adversarial examples (generated via techniques like FGSM) alongside regular inputs, allowing the model to adapt and learn discriminative patterns that reduce susceptibility to attacks. This ongoing exposure helps the model generalize better to unseen adversarial scenarios, particularly in complex visual tasks [8], [9], [11].

2.1. Data preparation and preprocessing

For our dataset, we utilized 20,000 images from the CelebA dataset [24] to train our proposed model. For preprocessing, first, to create steganographic images, we implemented a function that generated random strings and embedded these strings into the images using the least significant bit (LSB) method. The dataset was then split into training, validation, and test sets, with an 80:20 ratio between the training and test datasets. Then, each image was resized to 64×64 pixels to ensure efficient processing. Afterward, we generated steganographic images. Then, the images were prepared using EfficientNet's "preprocess input" function to ensure compatibility with the pre-trained model. After experimenting with various EfficientNet models and evaluating their performances, the EfficientNetB0 model was selected for feature extraction, generating feature vectors for each image.

Our steganalysis system consists of two main components: EfficientNet for feature extraction and a dense neural network for classification. The proposed model is designed using the sequential API from TensorFlow's Keras library, featuring multiple fully connected (dense) layers with enhancements for training efficiency and performance. It begins with an input layer tailored to the dimensionality of the training features. The first hidden layer contains 1024 neurons, followed by batch normalization for input stabilization, leaky rectified linear unit (ReLU) activation with a negative slope of 0.01 to address the dying ReLU problem, and a dropout layer with a 0.5 rate to prevent overfitting. Subsequent layers include 512, 256, 128, 64, 32, 16, 8, and 4 neurons, each employing batch normalization and leaky ReLU activation, with the first three also incorporating dropout. Leaky ReLU activation function and batch normalization [25] is also utilized for improved training [26]. The output layer features a single neuron with a sigmoid activation function for binary classification. The model is compiled with the Adam optimizer [27] (learning rate of 0.0003), BinaryCrossentropy loss function, and accuracy metric for performance evaluation. This architecture leverages batch normalization for better convergence, leaky ReLU to avoid dead neurons, and dropout to enhance generalization, making it robust and effective for the task at hand. Figure 1 shows the architecture of the model. The figure is generated using visualkeras [28].

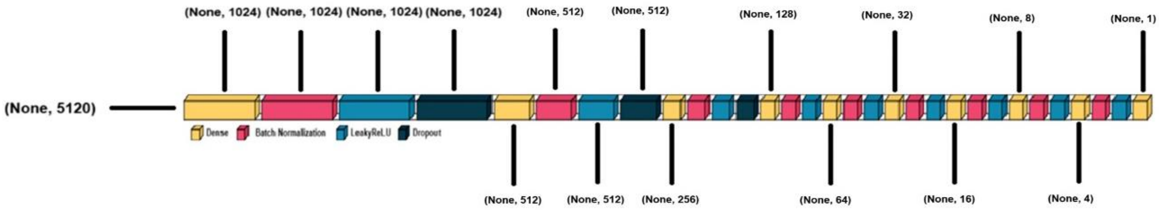

Figure 1. Model architecture

We employed the FGSM with a perturbation magnitude (epsilon) set to 0.01 to generate adversarial examples. These examples were used to augment the training data, enhancing the model's robustness to adversarial attacks. The adversarial training loop involved generating adversarial examples on the fly and combining them with original examples during each training epoch. This loop was designed to generate adversarial examples dynamically during each training epoch. That involved continuously producing new adversarial instances that were then mixed with the original training data. This approach not only increases the diversity of training examples but also helps the model develop a more nuanced understanding of the underlying data distribution, thus improving its overall resilience to adversarial perturbations.

The decision to apply the FGSM attack on the features extracted by EfficientNet, rather than directly on the original images, is driven by several key considerations:

- High-level feature manipulation: EfficientNet is designed to extract high-level, abstract features from images that are more representative of the underlying patterns and structures relevant for classification tasks. By applying FGSM on these high-level features, we ensure that the adversarial perturbations are crafted to target the most critical aspects of the data used by the classification model. This approach is likely to generate more effective adversarial examples that can challenge the model's robustness more directly.
- Dimensionality reduction: the feature space produced by EfficientNet is typically of lower dimensionality compared to the original image space. Operating in this reduced feature space allows for more efficient computation of adversarial examples. The FGSM attack, which involves calculating gradients, can be computationally intensive; thus, applying it in the feature space can significantly reduce the computational burden and expedite the generation of adversarial examples.

3. RESULTS AND DISCUSSION

To evaluate our proposed model system, we designed two scenarios. In the first scenario, the train set and test set have no adversarial instances. The purpose in scenario 1 is to evaluate the general performance of our steganalysis model against normal steganography images. Conversely, in the second scenario, we add adversarial instances to our train set on-the-fly. The evaluation also takes place on the test that contains adversarial instances. We trained three models: random forest model, base model, and adversarial model.

3.1. Scenario 1

In the first scenario, the dataset did not contain any adversarial instances. Additionally, we trained a random forest model for comparison with our proposed architecture. The results of this test are presented in Table 1.

Table 1. Performance comparison of models in scenario 1

Metric	Base model	Random forest
Accuracy	0.81450	0.65075
Precision	0.84255	0.64952
Recall	0.77195	0.64952
F1-score	0.80570	0.64952
ROC-AUC	0.81435	0.65074

3.2. Scenario 2

For the second part of the test, we included adversarial instances in the dataset. The adversarial model, which contains dropout and batch normalization layers, was trained with these adversarial examples. Figure 2 depicts the accuracy and loss plots for the adversarial model. In contrast, the base model and random forest did not encounter adversarial examples during their training. The results of this test are presented in Table 2.

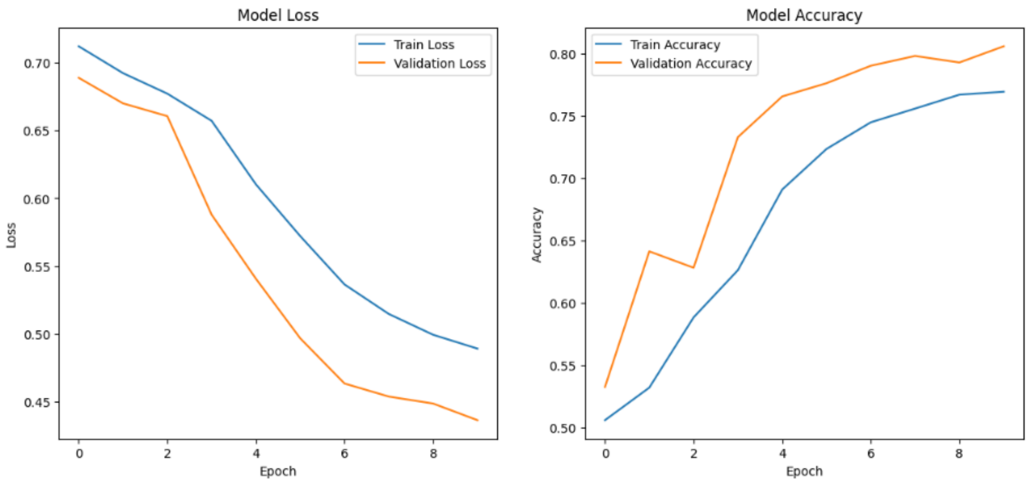

Figure 2. Training plot of the adversarial model

The evaluation of our proposed model system was conducted using two distinct datasets. The first dataset did not contain any adversarial instances, and we used it to benchmark the performance of our proposed model against a random forest model. As shown in Table 1, our proposed model significantly outperformed the random forest across all metrics. Specifically, our model achieved an accuracy of 0.81450, precision of 0.84255, recall of 0.77195, F1-score of 0.80570, and ROC-AUC of 0.81435, while the random forest lagged behind with an accuracy of 0.65075, precision of 0.64952, recall of 0.64952, F1-score of 0.64952, and ROC-AUC of 0.65074.

In the second part of the test, adversarial instances were included in the dataset to further evaluate the robustness of our proposed model. Unlike the base model and random forest, our adversarial model was trained with these adversarial examples and featured dropout and batch normalization layers. The results, presented in Table 2, illustrate a significant performance drop for the base model and random forest when faced with adversarial instances, with accuracies of 0.65400 and 0.54862, respectively. Their other metrics also declined substantially.

Table 2. Performance comparison of different models in scenario 2

Metric	Base model	Random forest	Adversarial model
Accuracy	0.65400	0.54862	0.72087
Precision	0.62016	0.61829	0.70983
Recall	0.78850	0.24586	0.74385
F1-score	0.69427	0.35182	0.72644
ROC-AUC	0.65446	0.54756	0.72095

Conversely, our adversarial model demonstrated superior resilience to adversarial attacks, achieving an accuracy of 0.72087, precision of 0.70983, recall of 0.74385, F1-score of 0.72644, and ROC-AUC of 0.72095. These results indicate that our adversarial training approach and the incorporation of dropout and batch normalization layers effectively enhanced the model’s robustness, maintaining higher performance levels in the presence of adversarial examples compared to models not trained with such examples. This underscores the importance of adversarial training in developing robust machine learning models capable of resisting adversarial perturbations.

The key difference between our proposed model and the random forest model lies in their architecture and adaptability to complex data patterns, particularly when faced with adversarial instances. Our proposed model is a neural network that incorporates advanced techniques such as batch normalization and dropout layers, allowing it to learn more intricate features from the data and maintain robust performance under varied conditions. As shown in the results from Table 1, this design led to a clear performance advantage over the random forest model, which relies on ensemble learning but lacks the deep learning model’s ability to capture non-linear relationships effectively. The random forest model exhibited significantly lower accuracy, precision, recall, F1-score, and ROC-AUC compared to our proposed model, highlighting the limitations of traditional machine learning methods in handling complex datasets without adversarial examples.

The inclusion of adversarial examples during training further underscores the importance of exposing models to diverse data variations to enhance robustness. Our adversarial model, designed with the capability to train on adversarial examples, showed remarkable resilience against adversarial attacks, as evidenced by its superior performance metrics in Table 2. In contrast, both the base model and the random forest experienced a notable decline in accuracy and other performance metrics when tested against adversarial examples, demonstrating their vulnerability. The adversarial training process, coupled with the architectural enhancements of our model, proved effective in improving its adaptability and robustness, maintaining higher performance levels compared to models that were not trained with adversarial examples. This higher performance is noticeable in Figure 3 where different metrics are compared across models.

EfficientNet has demonstrated exceptional capability as a feature extractor, capturing intricate details that contribute to precise classification in our steganalysis system. Its architectural efficiency allows it to focus on crucial patterns within the data, enabling a robust identification of features necessary for detecting hidden messages. Our results show that training with adversarial examples significantly enhances the robustness of the system, as reflected in the improved performance metrics under adversarial conditions. This approach effectively bolsters the model’s resistance to adversarial attacks, which are designed to mislead the model by introducing subtle perturbations. The base model, despite its initial effectiveness, revealed a susceptibility to these adversarial manipulations, underscoring the necessity for adversarial training in reinforcing model resilience.

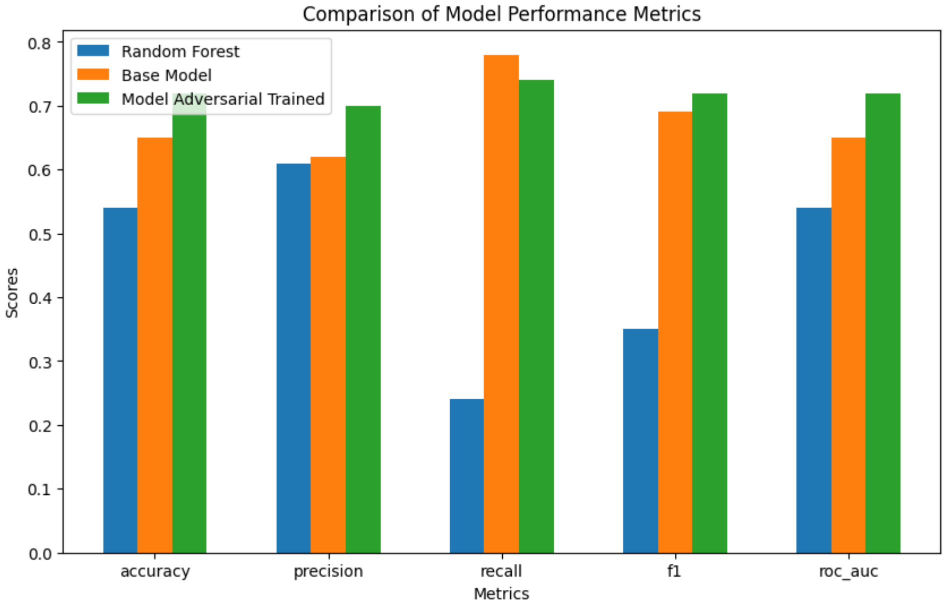

Figure 3. Metrics comparison between trained models

The random forest model, while simpler and providing a competitive baseline, lacks the robustness and adaptability offered by deep learning models l. Although random forests are advantageous in scenarios with limited computational resources and offer quick, straightforward implementation, they do not match the capability of deep learning models in handling complex data transformations and adversarial challenges. The combination of EfficientNet and adversarial training emerges as a powerful strategy for steganalysis, enabling the detection of hidden messages while maintaining robustness against adversarial perturbations. This highlights the critical importance of adversarial training in developing machine learning models that are not only accurate but also resilient to manipulative perturbations. Such robustness ensures reliability in practical applications where adversarial attacks pose a significant threat, reinforcing the model’s effectiveness and dependability in real-world scenarios. By incorporating these elements, our approach not only improves detection capabilities but also sets a foundation for future enhancements in machine learning security.

3.3. Comparison

Given the novelty of the proposed approach, direct comparison with existing methods is challenging due to the lack of directly comparable studies. However, to provide context for the performance of our method, we have selected a baseline approach and also offer a theoretical comparison to highlight the unique aspects of our work. Table 3 summarizes key differences and the potential advantages of our method.

Table 3. Comparison of methods

Methodology	Technique	Image size	Notable characteristics
Proposed method	DNN with adversarial training	64×64	Complex in training, high precision and accuracy, robust against adversarial examples
Random forest	Ensemble	64×64	Simple, vulnerable against adversarial examples, low accuracy and precision in comparison with other models.
Base model	DNN	64×64	Simple training, high accuracy in detecting normal examples, vulnerable against adversarial examples

3.4. Limitations

In this study, one of the main limitations is the lack of well-established baseline models for comparison. Given that this field is relatively new and rapidly evolving, there are few, if any, benchmark methods with standardized datasets or performance metrics that would provide a direct comparison for our model. Consequently, this limits our ability to quantitatively assess our model’s performance against a diverse set of approaches. Future work in this area will likely benefit from the development of more established benchmarks, which would provide clearer standards for evaluating and comparing advancements across different models.

4. CONCLUSION

This paper presents a novel steganalysis system that leverages EfficientNet for efficient feature extraction and adversarial training to enhance robustness. Our results demonstrate that the proposed system achieves high accuracy in detecting steganographic content and maintains resilience against adversarial attacks. This dual approach represents a significant advancement in steganalysis, offering a robust solution for secure information transmission in the real scenarios where the high performance of detection system is vital. Future research could explore the integration of advanced generative models and adversarial techniques to further improve the system's performance and robustness.

REFERENCES

[1] N. F. Johnson and S. Jajodia, "Exploring steganography: seeing the unseen," *Computer*, vol. 31, no. 2, pp. 26-34, Feb. 1998, doi: 10.1109/MC.1998.4655281.

[2] N. Meghanathan and L. Nayak, "Steganalysis algorithms for detecting the hidden information in image, audio and video cover media," *International Journal of Network Security & Its Application (IJNSA)*, vol. 2, no. 1, pp. 43-55, 2010.

[3] V. Verma, S. K. Muttoo, and V. B. Singh, "Detecting stegomalware: malicious image steganography and its intrusion in windows," in *Security, Privacy and Data Analytics: Select Proceedings of ISPDA* 2021, Singapore: Springer, 2022, pp. 103-116, doi: 10.1007/978-981-16-9089-1_9.

[4] Y. Shang, S. Jiang, D. Ye, and J. Huang, "Enhancing the security of deep learning steganography via adversarial examples," *Mathematics*, vol. 8, no. 9, pp. 1-10, 2020, doi: 10.3390/math8091446.

[5] K. Karampidis, E. Kavallieratou, and G. Papadourakis, "A review of image steganalysis techniques for digital forensics," *Journal of information security and applications*, vol. 40, pp. 217-235, 2018, doi: 10.1016/j.jisa.2018.04.005.

[6] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: a survey," *IEEE Access*, vol. 6, pp. 14410-14430, 2018, doi: 10.48550/arXiv.1801.00553.

[7] X. Yuan, P. He, Q. Zhu and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 9, pp. 2805-2824, Sept. 2019, doi: 10.1109/TNNLS.2018.2886017.

[8] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *ICLR 2017 Conference Track 5th International Conference on Learning Representations*, 2017.

[9] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: attacks and defenses," *arXiv*, 2017, doi: 10.48550/arXiv.1705.07204.

[10] M. Tan and Q. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," *Proceedings of the 36th International Conference on Machine Learning*, 2019, vol 97, pp. 6105-6114, doi: 10.48550/arXiv.1905.11946.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv*, 2014, doi: 10.48550/arXiv.1412.6572.

[12] S. M. Moosavidezfooli, A. Fawzi, and P. Frossard, "DeepFool: a simple and accurate method to fool deep neural networks," *Computer Vision and Pattern Recognition*, pp. 2574–2582, 2015, doi: 10.48550/arXiv.1511.04599.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *Security and Privacy*, 2017, doi: 10.1109/SP.2017.49.

[14] X. Qin, B. Li, S. Tan, W. Tang, and J. Huang, "Gradually enhanced adversarial perturbations on color pixel vectors for image steganography," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5110-5123, Aug. 2022, doi: 10.1109/TCSVT.2022.3148406.

[15] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, 2018, pp. 1-5, doi: 10.48550/arXiv.1704.02654.

[16] J. Liu, W. Zhang, Y. Zhang, D. Hou, Y. Liu, H. Zha, and N. Yu, "Detection based defense against adversarial examples from the steganalysis point of view," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4825-4834, doi: 10.48550/arXiv.1806.09186.

[17] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: protecting and vaccinating deep learning with jpeg compression," *arXiv*, 2017, doi: 10.48550/arXiv.1705.02900.

[18] N. Papernot, P. McDaniel, X. Wu, S. Jha and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, 2016, pp. 582-597, doi: 10.1109/SP.2016.41.

[19] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples," in *Proceedings of the 35 th International Conference on Machine Learning*, Stockholm, Sweden, PMLR 80, 2018.

[20] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," *CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 135-147, doi: 10.1145/3133956.3134057.

[21] O. Taran, S. Rezaeifar, T. Holotyak, and S. Voloshynovskiy, "Machine learning through cryptographic glasses: combating adversarial attacks by key-based diversified aggregation," *EURASIP journal on information security*, 2020, pp. 1-18, doi: 10.1186/s13635-020-00106-x.

[22] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[23] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*, 2019, pp. 7472-7482, doi: 10.48550/arXiv.1901.08573.

[24] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3730-3738, doi: 10.1109/ICCV.2015.425.

[25] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448-456, doi: 10.48550/arXiv.1502.03167.

[26] A. K. Dubey, and V. Jain, "Comparative study of convolution neural network's ReLU and Leaky-ReLU activation functions," in *Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018*, 2019, pp. 873-880, doi: 10.1007/978-981-13-6772-4_76.

[27] Kingma, P. Diederik, and J. Ba, "Adam: a method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2014, doi: 10.48550/arXiv.1412.6980

[28] P. Gavrikov, “visualkeras [Computer software],” GitHub repository, 2020, [Online]. Available: https://github.com/paulgavrikov/visualkeras (Accessed: Nov 20, 2024).

BIOGRAPHIES OF AUTHORS

Thakwan Akram Jawad M.S. master of science in Computer Engineering from Northern Cyprus (EMU) in Türkiye, Assistant Lecturer at Nineveh University, Department of System and Control Engineering College of Electronic Engineering Ninevah University Mosul 41002, Iraq. He can be contacted at email: thakwan.jawad@uoninevah.edu.iq.

Jamshid Bagherzadeh Mohasefi holds a Ph.D. in Computer Science from Indian Institute of Technology, India. He is currently Professor of the Department of Computer Science at the Urmia University. His research topics include artificial intelligence, machine learning, and network security. His research has been funded by the Urmia University, Iranian Telecom Ministry, and Iranian Ministry of Science, Research, and Technology. He can be contacted at email: j.bagherzadeh@urmia.ac.ir.

Mohammed Salah Reda Abdelghany holds a Ph.D. in Computer Science from the Faculty of Science at Damietta University, Egypt. He also earned a Master’s degree and a Pre-Master Diploma in Computer Science from the Faculty of Science at Minia University. Currently, he is affiliated with the Faculty of Computers and Artificial Intelligence at Beni-Suef university. His research interests span across various aspects of computer science and artificial intelligence. He can be contacted at email: msr_cs_87@du.edu.eg.