# Securing AI Models Against Backdoor Attacks: A Novel Approach Using Image Steganography

*Candra Ahmadi*, Jiann-Liang Chen, Yu-Ting Lin*

*Department of Electrical Engineering, National Taiwan University of Science and Technology, Taiwan*
*D11007809@mail.ntust.edu.tw, lchen@mail.ntust.edu.tw, m11007501@gapps.ntust.edu.tw*

## Abstract

Artificial Intelligence (AI) has become ubiquitous, transforming numerous domains including traffic sign recognition, defect detection, and healthcare. However, this widespread adoption has brought about significant cybersecurity challenges, particularly in the form of backdoor attacks, which manipulate training datasets to compromise model integrity. While the integration of AI has proven beneficial, there is a lack of comprehensive strategies to protect AI models from these covert attacks, necessitating innovative approaches for securing AI systems. In this study, we demonstrate a novel methodology that integrates image steganography with deep learning techniques, aiming to obscure backdoor triggers and enhance the resilience of AI models against these attacks. We employ a diverse set of AI models and conduct extensive evaluations in a traffic sign recognition scenario, specifically targeting the STOP sign. The results reveal that shallow models are challenged in learning trigger information and are sensitive to trigger settings, while deeper models achieve an impressive 98.03% attack success rate. The image steganography technique used requires minimal data adjustments, making the triggers more challenging to detect than with traditional methods. Our findings underscore the stealth and severity of backdoor attacks, emphasizing the need for advanced security measures in AI and contributing to the broader understanding and development of robust protections against such attacks.

**Keywords:** Artificial Intelligence security, Backdoor attack, Deep learning, Image recognition, Image steganography

## 1 Introduction

Artificial Intelligence (AI) has emerged as a transformative force, significantly impacting various domains such as traffic sign recognition, defect detection, unmanned stores, and healthcare [1]. This technological revolution has led to a rapid and extensive integration of AI models, particularly in image recognition, into societal structures and industry practices.

The broad adoption of AI has unlocked unprecedented advancements, enhancing efficiency, accuracy, and innovation across sectors. However, the expansive deployment of AI technologies has concurrently introduced new cybersecurity challenges, necessitating an urgent and focused examination of AI model security. Backdoor attacks on image recognition models exemplify a critical threat, posing substantial risks across diverse applications [2-7].

Backdoor attacks, a prominent concern in AI security, entail covert manipulations of the training dataset to embed vulnerabilities into the resultant model. These compromised models exhibit dual behavior, functioning normally in regular scenarios but misclassifying in a targeted manner when a specific trigger is present during inference [4-6]. The implications of such attacks are profound, affecting various sectors and individuals reliant on AI systems.

Although there have been notable strides in AI security, existing countermeasures against backdoor attacks reveal limitations, particularly in their ability to detect and mitigate these threats effectively. Current approaches have not sufficiently addressed the challenge of camouflaging attack triggers, leaving them potentially detectable by adversaries [4, 6].

This research introduces a novel method that synergistically combines image steganography with advanced deep learning techniques to obfuscate backdoor attack triggers. By embedding the triggers within digital images in a concealed manner, image steganography enhances the difficulty of detecting these triggers, thereby bolstering AI model security [8-9].

To empirically validate the efficacy of the proposed solution, this paper conducts an extensive simulation involving backdoor attacks on a traffic sign image recognition model. The study encompasses seven trigger-setting conditions and employs three distinct deep learning models, ensuring a thorough and comprehensive evaluation [10-13]. The findings underscore the urgent need for robust AI model security and the effectiveness of the proposed methodology.

This work contributes significantly to the existing literature on backdoor attacks, enhancing the understanding of these threats, and introducing innovative strategies for concealing attack triggers. The study emphasizes the imperative for stringent AI model security and offers valuable recommendations for mitigating backdoor attacks, guiding researchers, cybersecurity experts, and AI practitioners towards developing more secure AI systems [14-16].

As AI continues to evolve and infiltrate various sectors, the insights and recommendations from this study gain

*Corresponding Author: Candra Ahmadi; E-mail: D11007809@mail.ntust.edu.tw
DOI: 10.53106/160792642024052503012

paramount importance in safeguarding AI-dependent societies against backdoor attacks. The strategic application of image steganography, coupled with informed AI model layer selection, becomes crucial in this context.

The subsequent sections of this manuscript delve deeper into the intricacies of image recognition models, the vulnerabilities in AI, backdoor attacks, and the requisite for robust security mechanisms. Readers will be provided with a detailed exploration of the subject, accompanied by empirical evidence and methodological discussions. The paper is structured to facilitate a seamless flow of information, starting with a comprehensive literature review AI and Security Measures in Section II, followed by the proposed system in Section III, results and discussion in Section IV, and concluding with recommendations and future directions in Section V.

AI solidifies its role in critical decision-making and societal functions, ensuring the integrity and security of its models becomes indispensable. This research stands as a foundational step towards a future where trust in AI is validated and secured, paving the way for innovative and resilient AI security solutions.

# 2 AI and Security Measures

## 2.1 Advances in AI Image Recognition

Artificial intelligence image recognition technology allows computers to use artificial intelligence models to identify images. In 1998, Y. Lecun et al. proposed LeNet [17] used convolution, pooling, and fully connected layer architecture for the first time in the font image recognition and classification task to achieve a good prediction level. It is also one of the earliest convolutional neural networks. Its model architecture is shown in Figure 1. By introducing convolutional layers instead of fully connected layers, the number of model parameters is reduced but still comparable to that of support vector machines, making convolutional neural networks one of the mainstream algorithms in image recognition.

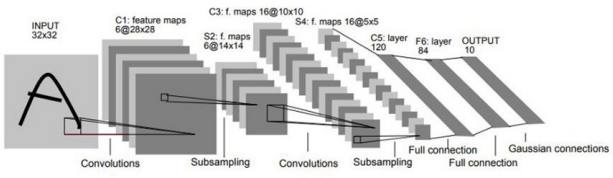


**Figure 1.** LeNet Architecture [17]

In 2012, A. Krizhevsky et al. proposed AlexNet [18] made a significant breakthrough in image recognition by deepening the number of model layers, changing the nonlinear activation function, and using techniques such as dropout to avoid model overfitting. Demonstrate that increasing the convolutional neural network layers can bring benefits and that the feature information learned by the model can exceed the number of human-selected features. Won the ImageNet recognition challenge with a considerable lead, making the convolutional neural network as a mainstream architecture.

In 2014, K. Simonyan et al. proposed very deep convolutional networks [19] through data-oriented multiple scale training, reducing the size of the convolution kernel on the model architecture and superimposing them, and importing them into the architecture Max pooling layer. This architecture enables convolutional neural networks to reach a deeper model layer and achieve better prediction accuracy. The study confirmed that the deep neural network using small convolution kernels is better than the shallow neural network with large convolution kernels, and can use multi-layer small convolution kernels to achieve the same receptive field as large convolution kernels.

With the development of convolutional neural networks, the deeper the model architecture, how to add more model layers to improve the model's performance becomes extremely important. In 2016, K. He et al. studied that deep convolutional neural networks are difficult to achieve reasonable accuracy [20]. To solve this problem, they believe each model layer should include the previous function as one of the inputs. Based on this, it is proposed to add a residual block to the convolutional neural network architecture to break through the gradient disappearance problem that will occur when the convolutional neural network updates the model weight parameters in the deep architecture. At the same time, it is proved that using a residual network can achieve a better network optimization effect and construct a deeper neural network structure to improve the accuracy of the model, which will have a profound impact on the subsequent deep neural network design.

## 2.2 Ensuring the Security of AI Models

This section will discuss the artificial intelligence model security, including the description of adversarial attacks that target artificial intelligence and the MITER ATLAS framework that targets artificial intelligence model security. Artificial intelligence model deployment can be divided into two blocks: model training and model inference. The model training stage is when the artificial intelligence model uses its neural network to learn and complete a particular job, such as traffic sign and face recognition. The model inference stage deploys the artificial intelligence model after model training to the field for actual use.

With the proliferation of network attacks and the widespread application of artificial intelligence, attacks against the artificial intelligence model have emerged. This attack is called an adversarial attack, including Model Evasion, Functional Extraction, Model Poisoning, and Model Inversion. Attack stages and descriptions of adversarial attacks can see at [21-22]. The purpose of the backdoor attack is to embed a hidden backdoor into the deep neural network. This will make the attacked artificial intelligence model normally perform on benign samples. But when the input data contains triggers defined by the attacker, the backdoor will be triggered, making the prediction results maliciously changed and giving targeted wrong results. In [21] provide lists the detailed descriptions of the expected attack behaviors involved in this research according to the MITER ATLAS framework, including Reconnaissance and Resource Development before the attack, Persistence and ML

Attack Staging during the attack, and Impact generated after the attack.

### 2.3 Backdoor Attack in Deep Learning

The purpose of this section is to review the current state of research on backdoor attacks in deep learning. Deep neural networks have been deployed and applied in various fields, and triggers are added to the training data set in the model training phase, resulting in changes during the model training period so that the trained model produces targeted wrong predictions in the inference phase. This kind of attack is called a backdoor attack. It is subdivided into Corrupted-label attacks and Clean-label attacks according to whether it involves label manipulation during the interference training data set.

Gu et al. first demonstrated the possibility of injecting backdoors into DNNs in 2017 [23]. The study made the trained model generate backdoors by injecting backdoor triggers in the traffic signs and handwritten digit datasets. Inputs with triggers are subsequently predicted as targeted error classes.

In 2021, Wenger et al. tried to carry out backdoor attacks in the physical world through self-collected facial recognition data sets. They are bringing backdoor triggers from the data level into real life by using common objects, such as stickers, earrings, etc. as triggers, proving the possibility of backdoor attack risks in the physical world [24].

A study on Clean-label attacks was published by Shafahi et al. the following year [25]. This study achieved a backdoor attack without changing the labels of the training datasets. This study confirms that there is still the danger of backdoor attacks when only the eigenvalues of the training datasets are interfered with.

Yao et al. studied the backdoor attack in transfer learning in 2019 and tested it on traffic signs, handwritten numbers, and face recognition datasets [26]. The simulation results achieved a high attack success rate. The same year, Chen et al. used Activation Clustering to detect backdoor attacks [27]. When the trigger setting is clearly different from the normal input, it can achieve good detection ability.

Research on backdoor attacks has sprung up like mushrooms after rain in recent years. Including the implantation of backdoor triggers in the model retraining stage proposed by Costales et al. in 2020 to carry out backdoor attacks in datasets such as handwritten numbers and self-driving cars [28]. Lin et al. utilize normal objects as backdoor trigger settings to conceal backdoor attacks [29]. And an attack method that can set multiple backdoors at the same time was also proposed by Zhong et al. in the same year and realized in datasets such as traffic signs [30].

At the same time, there is also research on backdoor attacks for federated learning, which was proposed by Bagdasaryan et al. in 2020 [31]. Through the model-poisoning attack in the model aggregation stage, the global model is endangered by backdoor attacks. And the backdoor attack scheme was studied by Quiring et al. by manipulating image zoom so that the backdoor trigger will only appear at a specific zoom ratio [32]. Both of the above two studies were implemented in data sets such as CIFAR-10.

In 2020, Pang et al. cut into the research from another angle by analyzing the interaction between normal samples and backdoor samples for the model and by simulating in datasets, such as traffic signs and CIFAR-10, to replace backdoor attack defense measures [33]. Expand another unique point of view. Rakin et al. achieved backdoor attacks through a small number of neuron-flipping techniques in the CIFAR-10 and ImageNet datasets [34].

Zeng et al. focus on the research of trigger setting, explore the impact of trigger setting in different regions of the image, and realize backdoor attack in CIFAR-10 by adopting trigger setting in different spatial positions [35]. Bagdasaryan et al. tried using code poisoning techniques to carry out backdoor attacks in multi-task learning tasks, experimenting with ImageNet and handwritten digital datasets [36].

Severi et al. extended the object of backdoor attacks to target malware recognition models. Using EMBER, Contagio and Drebin datasets for simulation, further information changes are made to achieve attack behavior by analyzing the degree of influence of features [37]. Li et al. designed an autoencoder for trigger settings to hide triggers and simulated backdoor attacks on datasets such as ImageNet and Ms-Celeb [38].

In 2022, Salem et al. proposed a method of randomizing trigger settings for datasets such as handwritten numbers and face recognition and further completed the backdoor generation network and conditional backdoor generation network to develop backdoor attacks [39]. Tian et al. tried a new backdoor attack on the traffic sign and CIFAR-10 datasets in the same year, specifically for the backdoor attack on the lightweight model stage. The backdoor will not appear until the original model is lightweight [40].

### 2.4 Steganography

Steganography is a covert communication technique that hides information such as documents, messages, or images so that the hidden information is invisible to observers. Unlike cryptography, which transforms data into another form of protection, steganography is the practice of hiding information within another. Take image steganography as an example. Specifically, it is to hide the information to be hidden into the normal image through algorithm design and then extract the hidden information from the image through algorithm design.

For image steganography, the common practice is the least significant bit (LSB) substitution [41]. The principle is that the lower bit information in the image item will not have a significant color change to the human eye. By replacing the LSB information, you can put the information you want to hide and extract it when needed.

Another steganography technique works by reading the pixel values of two images in binary and combining their most significant bits (MSB). Since MSB contains more image information than LSB, this method can largely preserve the features of the image to be hidden [42].

## 3 Method

### 3.1 System Architecture

This study presents a system comprising five stages: data

collection, trigger settings, data preprocessing, model training and tuning, and model prediction and evaluation as shown at Figure 2. The German Traffic Sign Recognition Benchmark (GTSRB) dataset is used, encompassing 43 different traffic sign labels. Triggers are then set up and added to images, followed by data preprocessing which includes normalization, segmentation, and augmentation. Three deep neural network models, each representing different architectural depths, are then trained and tuned. The impact and effectiveness of backdoor attacks on these models under different scenarios are evaluated in the final stage.

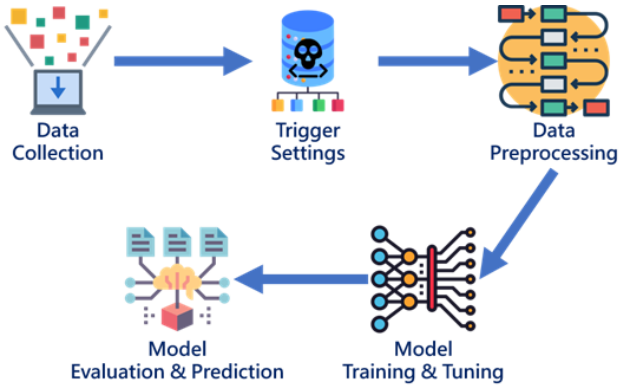


**Figure 2**. System architecture

### 3.2 Data Collection

This research uses the German Traffic Sign Recognition Benchmark (GTSRB) dataset from Kaggle, which consists of over 50,000 images across 43 distinct traffic sign categories, to simulate and analyze backdoor attacks on an image recognition model. Table 1 show GTSRB Dataset Overview for this research. The dataset's variety in image sizes, lighting and weather conditions, rotations, and occlusions accurately represent real-world scenarios, making it an ideal choice for studying real-world applicable traffic signs (Figure 3) [43-44].

**Figure 3**. GTSRB dataset schematic

**Table 1.** GTSRB dataset overview

| Dataset | Object | Source | Image size | Number of categories | Amount of data |
|---|---|---|---|---|---|
| GTSRB | Traffic signs | Kaggle | 15x15 to 250x250 | 43 | 51,838 |

### 3.3 Trigger Settings

This study focuses on the strategic implementation of a threat model and backdoor triggers. The threat model conceptualizes a backdoor attack scenario in which an adversary alters the training dataset to misclassify future inputs with triggers. Adversaries may range from malevolent dataset suppliers to external model training service providers. The research emulates varying triggers and model architectures to assess backdoor attack risks. The impacts of image scale and trigger size on backdoor attacks are analyzed with two image scales (100 x 100 x 3 and 224 x 224 x 3 pixels) and seven trigger sizes. Image steganography improves trigger concealment, reserving the last three bits of an image for trigger settings. The backdoor trigger's application extends to the full image and label category modification, generating a dataset of clean and backdoor data. The traffic sign "STOP" at Figure 4 is used as the trigger image, aimed to misclassify input images with triggers. Subsequently, backdoor samples are created for further training and risk evaluation.


**Figure 4**. Trigger image schematic

### 3.4 Model Training & Tuning

This study compares the impact of backdoor attacks on This study evaluates the impact of backdoor attacks on different neural network models, namely shallow, middle, and deep layer models. The models are trained on clean datasets, then on datasets composed of clean and backdoor data to study the effect of network depth and trigger scales on backdoor attacks. The shallow layer model uses a configuration of four convolutional layers, two pooling layers, and two fully connected layers, with ReLU and Softmax activation functions and the Adam optimizer. The middle layer model employs the VGG-16 framework with 13 convolutional layers, interspersed with pooling layers and three fully connected layers. The deep layer model, based on ResNet-50, addresses network degradation through a residual network design comprising 16 residual blocks and a single convolutional layer. The study leverages these models to provide an insightful risk assessment of backdoor attacks across various network depths.

### 3.5 Model Evaluation & Prediction

This study revolves around the dual optimization problem of learning clean data and backdoor trigger features in a model's training process. The trained models are then evaluated using two key metrics - Attack Success Rate (ASR) and Clean Data Accuracy (CDA). They are tested on two datasets, the clean test dataset ($D_{clean\ test}$) and the backdoor test dataset ($D_{backdoor\ test}$). ASR measures the success rate of backdoor attacks on the model, observing if the input with triggers can be accurately predicted to the target label "STOP". On the other hand, CDA gauges whether learning trigger-related features impacts the prediction of clean data. Another metric, Test Accuracy Loss, is calculated as the difference between the CDA of the clean and backdoor models. Overall, this study explores how ASR and CDA metrics are impacted by seven different trigger settings across three models, with ASR results compared to previous studies.

## 4 Analysis Performance and Discussion

### 4.1 Shallow Layer Model

This section presents the experimental outcomes of the shallow layer model subjected to backdoor attacks, with metric details enumerated in Table 2. Post experimentation, the model yielded an accuracy of 96.96% on the clean dataset, as visualized in Figure 5, illustrating the accuracy and loss function curves during training along with the prediction results. The model displayed proficient classification capabilities for a majority of images, struggling only with darker, complex road signs. In backdoor attack trials, varying trigger configurations resulted in significant fluctuations in the Attack Success Rate (ASR) between 60.93% and 92.63%, with the Clean Data Accuracy (CDA) remaining more consistent, ranging from 93.84% to 96.53%. Figures 5 display confusion matrices for the experimental outcomes of backdoor and clean test datasets, respectively, indicating that the model can mispredict a small portion of clean data after trigger information learning, suggesting even simple network structures can be compromised with a well-designed trigger setup. In Figure 6 we can see backdoor test confusion matrix and clean test confusion matrix.

### 4.2 Middle Layer Model

The experimental results from the middle layer model's backdoor attack experimentation are presented herein, with specific evaluation metrics detailed in Table 3. Following experimental trials, the model achieved an accuracy of 97.07% on clean datasets. The progression of training accuracy, loss function curve, and test dataset predictions are captured in Figure 7. This model's superior performance, vis-a-vis the shallow layer model, becomes evident when it accurately predicts images that its predecessor failed to. Its ability to identify blurred and dark images is attributed to the enhancement in model parameters and architecture. Backdoor attack experiments revealed that certain middle layer models could reach an ASR of 90.06%~95.63% and a CDA of 96.43%~98.11% across seven trigger settings. Unlike the shallow layer model, variation in triggers and image resolution caused only a 5.57% change in ASR and halved the fluctuation in CDA for the middle layer model. Figure 8 respectively depict the confusion matrix for the backdoor and clean test datasets in the configuration with the highest ASR. The model achieved an ASR of 95.63% for trigger-containing samples, impacting the clean data less than the shallow layer model. However, as the network architecture deepens, the model becomes more sensitive to trigger settings and more vulnerable to backdoor attacks.

**Table 2**. Shallow layer model experimental results

| Triger size /Original image size | Clean Data | 50/100 | 60/100 | 75/100 | 100/100 | 100/224 | 150/224 | 200/224 |
|---|---|---|---|---|---|---|---|---|
| Clean data accuracy | 0.9696 | 0.9384 | 0.9629 | 0.9606 | 0.9594 | 0,9653 | 0.9552 | 0.9566 |
| Attack success rate | - | 0.6200 | 0.6093 | 0.6176 | 0.8133 | 0,9263 | 0.9210 | 0.9036 |
| Test accuracy loss | - | 0.0312 | 0.0067 | 0.009 | 0.0102 | 0,0043 | 0.0144 | 0.01300 |

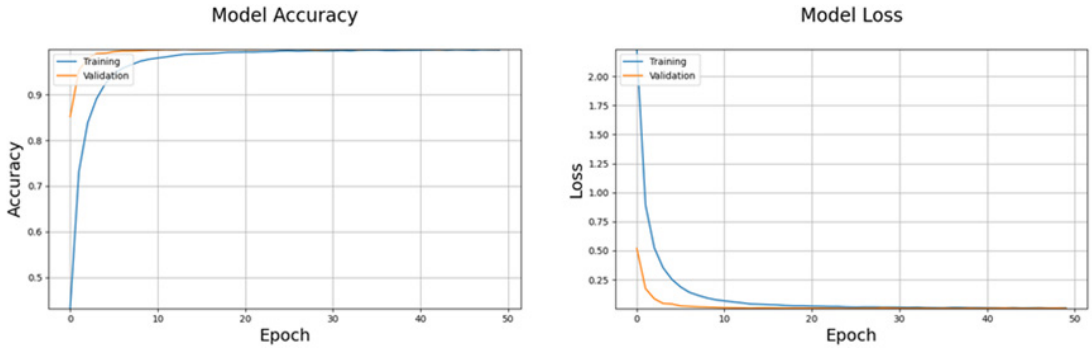


**Figure 5**. Shallow model training process diagram

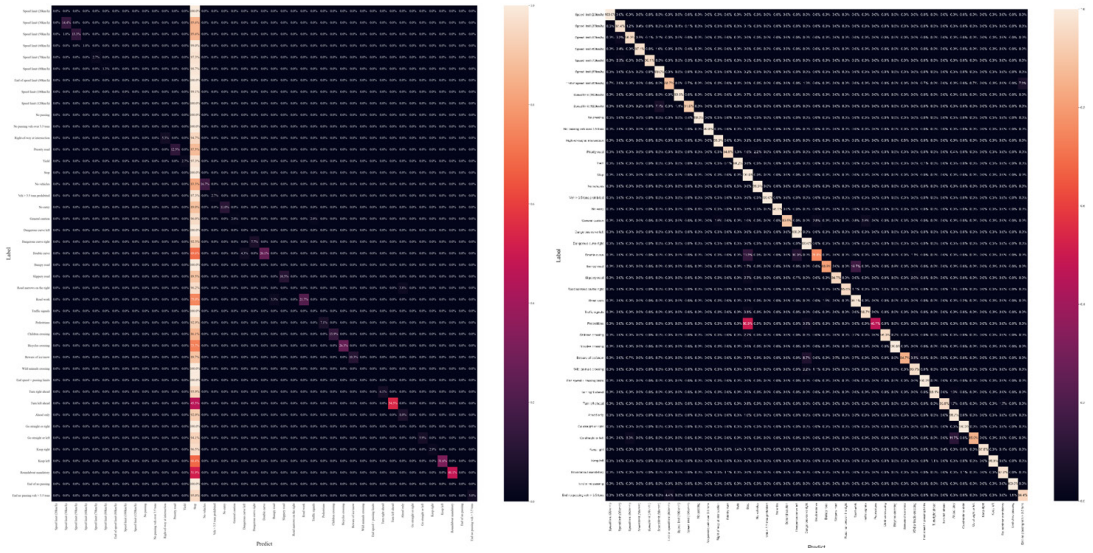

**Figure 6**. Backdoor test confusion matrix and clean test confusion matrix

**Table 3**. Shallow layer model experimental results

| Triger size /Original image size | Clean data | 50/100 | 60/100 | 75/100 | 100/100 | 100/224 | 150/224 | 200/224 |
|---|---|---|---|---|---|---|---|---|
| Clean data accuracy | 0.9707 | 0.9811 | 0.9761 | 0.9807 | 0.9671 | 0,9735 | 0.9643 | 0.9798 |
| Attack success rate | - | 0.9060 | 0.9356 | 0.9253 | 0.9006 | 0,9563 | 0.9533 | 0.9386 |
| Test accuracy loss | - | -0.0104 | -0.0054 | -0.0100 | 0.0036 | -0.0028 | 0.0064 | -0.0091 |

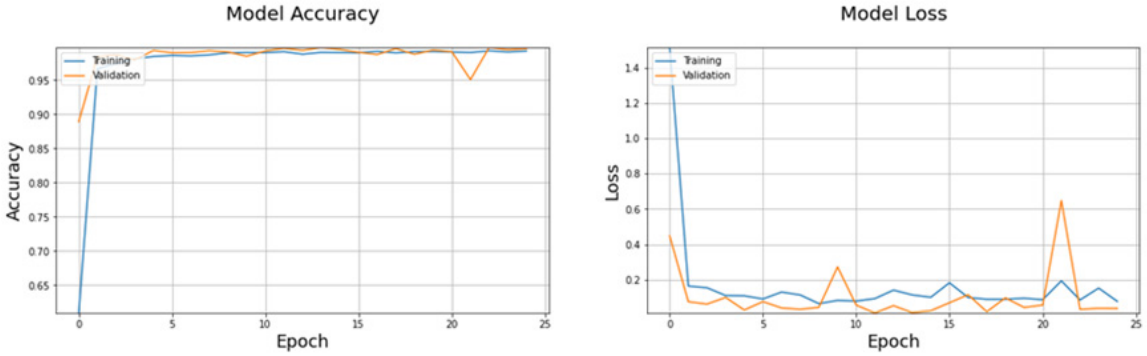


**Figure 7**. Middle layer model training

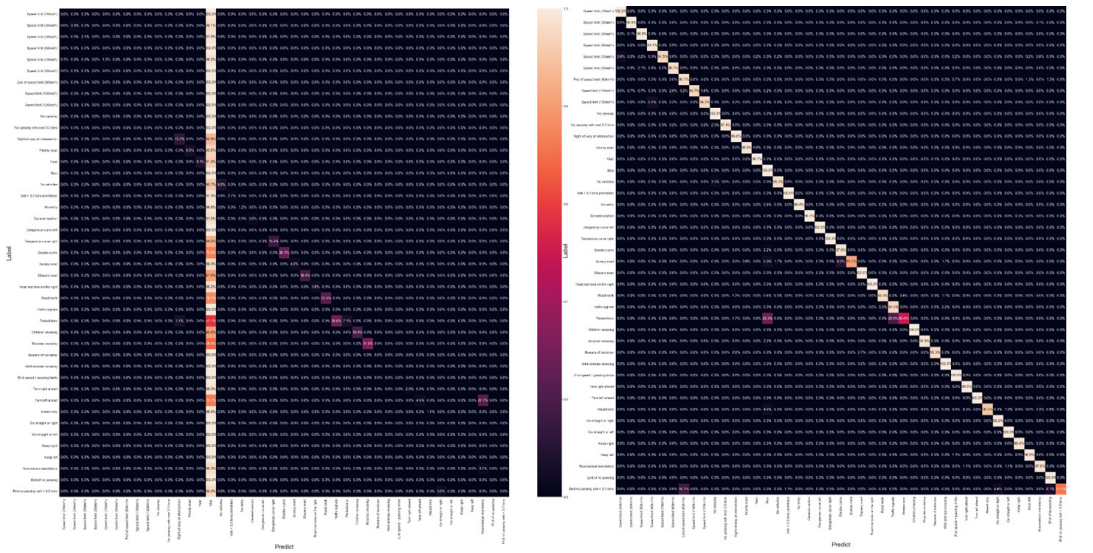

**Figure 8.** Backdoor test confusion matrix for middle layer model and clean test confusion matrix for middle layer model

### 4.3 Deep Layer Model

In this section will present the experimental results of the deep layer model's performance in a backdoor attack scenario as we can see at Table 4, Figure 9 and Figure 10. The model achieved an accuracy of 98.18% on a clean dataset, demonstrating improved performance on more complex images. During the backdoor attack experiment, the model showed variable Attack Success Rate (ASR) between 88.3% and 98.03% and Clean Data Accuracy (CDA) between 95.15% and 96.5% across seven different trigger configurations. Compared to shallower models, the deep model was more sensitive to trigger settings, showing stronger learning of the hidden trigger information with minimal impact on clean data, thus making it more susceptible to potential backdoor attacks. Confusion matrices were constructed to illustrate these results for both clean and test datasets.

**Table 4**. Shallow layer model experimental results

| Triger size /Original image size | Clean data | 50/100 | 60/100 | 75/100 | 100/100 | 100/224 | 150/224 | 200/224 |
|---|---|---|---|---|---|---|---|---|
| Clean data accuracy | 0.9818 | 0.9515 | 0.9616 | 0.9567 | 0.9517 | 0,9650 | 0.9611 | 0.9574 |
| Attack success rate | - | 0.8830 | 0.9590 | 0.9540 | 0.9160 | 0,9793 | 0.9803 | 0.9753 |
| Test accuracy loss | - | 0.0303 | 0.0202 | 0.0251 | 0.0301 | 0.0168 | 0.0207 | 0.0244 |

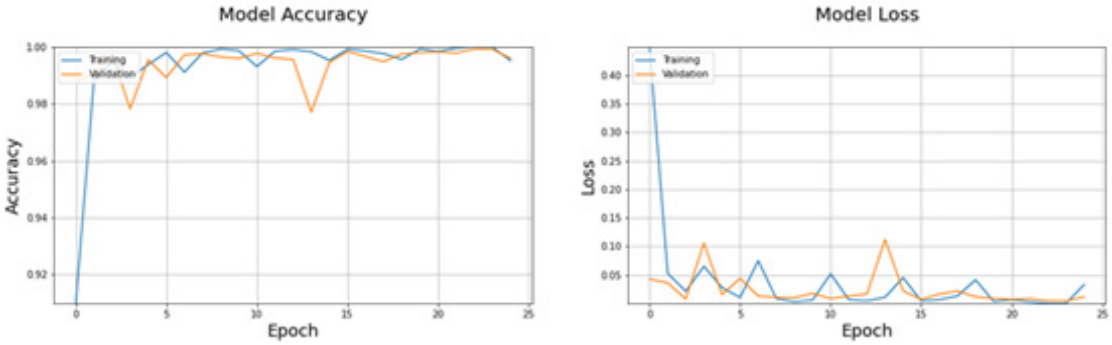


**Figure 9**. Deep layer model training process

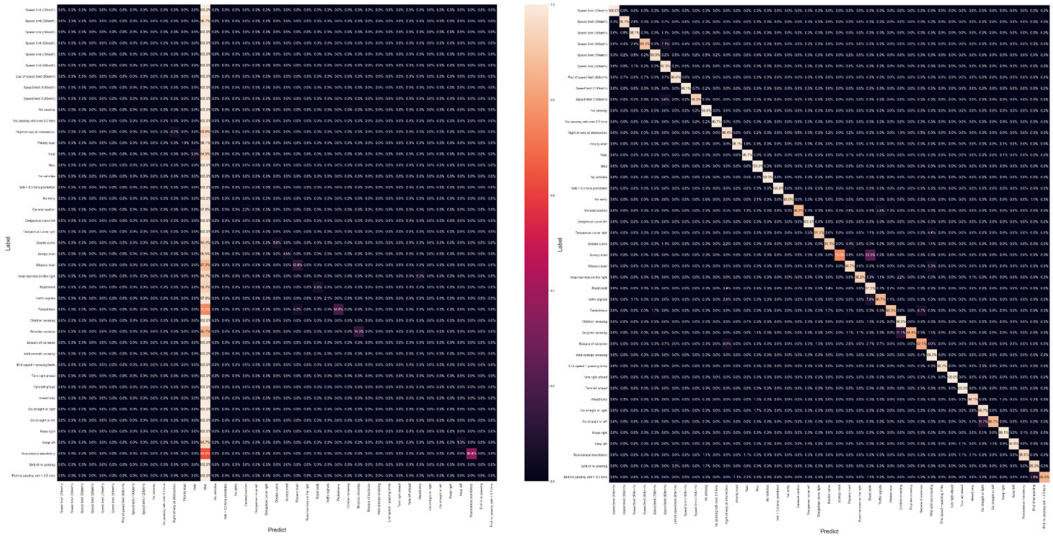

**Figure 10**. Deep model backdoor test data confusion matrix and shallow model clean test data confusion matrix

### 4.4 Backdoor Trigger

This study utilizes a backdoor trigger setup that is virtually imperceptible to the human eye. Figure 9 demonstrates the process of trigger setup, comparing the clean image, the trigger schematic, and the image post trigger addition, revealing a significant degree of concealment. The trigger's impact on the image's RGB channels is illustrated in Figure 10, further highlighting the stealthy nature of the trigger setup. Despite their low visibility, these triggers provide fixed features for model learning, thereby increasing the vulnerability of models to backdoor attacks.

### 4.5 Results Summary

This research analyzes the impact of backdoor attacks through simulations on three different model architectures under varied trigger configurations. Special attention is

given to the configuration of the triggering mechanism, as illustrated in Figure 11, which depicts the trigger setting schematic. Furthermore, the complexity of the system is enhanced by the introduction of a three-channel trigger setup, detailed in Figure 12, offering a more comprehensive understanding of the operational dynamics. The clean data accuracy (CDA) metric remains largely stable across shallow and medium-deep layer models, even when backdoor triggers are introduced, indicating their subtle impact. However, the attack success rate (ASR) varies, with shallow layer models showing sensitivity to trigger size changes in low-resolution images due to their limited learning capabilities. Conversely, middle and deep layer models display a strong fitting ability, achieving high ASR in low-resolution images, and their learning capacity grows with the increase in image resolution and complexity of trigger information, with the ASR reaching up to 98.03%.


**Figure 11**. Trigger setting schematic


**Figure 12**. Three-channel trigger setup schematic

### 4.6 Comparison with Other Studies

This section compares the experimental results of this study with those of previous studies.

Table 5 presents the model metrics of this study compared with previous similar studies using the same dataset or similar model architectures. The experiments were conducted using different depth convolutional neural networks and trigger setups with the dataset GTSRB and evaluated by keeping the test set data in advance. The evaluation metrics include Poison Rate, Test Accuracy Loss and ASR.

The first comparator is a backdoor attack by adding a perturbation mask to the image proposed by H. Zhong et al [30]. The second one is a backdoor attack by E. Wenger et al. that collects real face data in the physical world and sets up real objects as triggers [24]. The third one is A. Salem et al. used a backdoor generation network to combine the means of setting up a backdoor attack trigger with a generation countermeasure network to perform a backdoor attack [39]. Finally, Y. Tian et al. proposed a backdoor attack technique by activating backdoors in the process of model lightweight [40].

All of these approaches have their own advantages and disadvantages and their own means to hide the triggers. Still, some of them require a higher Poison Rate, meaning that the data with triggers need to have a more significant percentage in the training set, and some of them are more complicated and cumbersome in setting the triggers, or have room for growth in the final ASR metric.

Table 5 shows that the proposed method does not differ significantly from other studies in terms of test accuracy loss, and increases at least 2.9% in the evaluation metric ASR, and decreases at least 2.7% in the percentage of backdoor triggers set in the dataset. These metrics perform well compared to related studies, further confirming the feasibility and risk of backdoor attacks.

## 5 Conclusion

In conclusion, this study has successfully presented an innovative solution to address the critical issue of backdoor attacks in AI models, utilizing the synergistic combination of image steganography and deep learning techniques. By integrating image steganography, we have introduced a novel approach to obscure backdoor triggers, thereby enhancing the resilience of AI models against these insidious attacks. Our method has demonstrated its effectiveness in a traffic sign recognition scenario, providing a robust response to the challenges posed by backdoor attacks.

The concept of our proposed solution revolves around the strategic use of image steganography to conceal backdoor triggers within digital images, making them significantly more challenging for adversaries to detect. This approach not only fortifies AI models against covert manipulations but also maintains the integrity of the model's performance in clean data scenarios. Through extensive evaluations and diverse trigger-setting conditions, our solution has showcased its ability to effectively mitigate the risks associated with backdoor attacks, particularly in complex deep learning models.

The most notable result from our empirical studies is the varying degrees of susceptibility among different AI models, with deeper models exhibiting a 98.03% attack success rate, highlighting the stealth and severity of backdoor attacks. However, the implementation of our image steganography technique has proven to be a formidable countermeasure, requiring minimal data adjustments and thus ensuring the triggers remain inconspicuous.

**Table 5** Comparison with the results of different studies.

| Method | Backdoor embedding in convolutional neural network models via invisible perturbation (2020) | Backdoor attacks against deep learning systems in the physical world (2021) | Dynamic backdoor attacks against machine learning models (2022) | Stealthy backdoors as compression artifacts (2022) | Our proposed system |
|---|---|---|---|---|---|
| Algorithm | Multi-scale convolutional networks [39] | ResNet-50 | VGG-19 | VGG-16 | CNN |
| Dataset | GTSRB | Facial recognition dataset | CIFAR-10 | GTSRB | GTSRB |
| Poision rate | 4.7% | 83% | 10% | | Less than 2% |
| Test accuracy loss | 0.82% | 2% | 0.3% | 0.5% | 1.68% |
| ASR | 88.22% | 95% | 92.4% | 87.3% | 97.9% |

By re-emphasizing the performance claims, it is evident that our proposed methodology stands out in its ability to secure AI models from backdoor attacks, ensuring a robust and resilient AI security landscape. The application of image steganography, in conjunction with deep learning, marks a significant advancement in the field, setting a new benchmark for AI security measures.

As we move forward, the imperative for innovative and effective security measures becomes increasingly critical. This research lays the groundwork for future endeavors, guiding the way for the development of more secure and trustworthy AI systems. With the insights and recommendations provided in this study, the AI community is well-equipped to tackle the challenges of backdoor attacks, ensuring a secure and reliable future for AI applications across various domains.

## References


[1] A. Hemmati, A. M. Rahmani, The Internet of Autonomous Things applications: A taxonomy, technologies, and future directions, *Internet of Things*, Vol. 20, pp. 1-12, November, 2022.

[2] F. Kanakov, I. Prokhorov, Analysis and applicability of artificial intelligence technologies in the field of RPA software robots for automating business processes, *Procedia Computer Science*, Vol. 213, pp. 296-300, 2022.

[3] The MITRE Corporation, MITRE ATT&CK, Retrieved from https://attack.mitre.org/resources/faq/, last visited on 2023/05/23.

[4] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, R. Shokri, Enhanced membership inference attacks against machine learning models, *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, Los Angeles, CA, USA, 2022, pp. 3093-3106.

[5] The MITRE Corporation, MITRE ATLAS, Retrieved from https://atlas.mitre.org/, last visited on 2023/05/20.

[6] S. Goldwasser, M. P. Kim, V. Vaikuntanathan, O. Zamir, Planting undetectable backdoors in machine learning models, *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, Denver, CO, USA, 2022, pp. 931-942.

[7] A. Ali, Z. Farid, A. H. Al-kassem, Z. A. Khan, M. Qamer, K. F. K. Ghouri, M. A. Sakhnani, A. M. Momani, Development and use of Artificial Intelligence in the Defense Sector, *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, Dubai, United Arab Emirates, 2023, pp. 1-10.

[8] I. Ilahi, M. Usama, J. Qadir, M. U. Janjua, A. Al-Fuqaha, D. T. Hoang, D. Niyato, Challenges and countermeasures for adversarial attacks on deep reinforcement learning, *IEEE Transactions on Artificial Intelligence*, Vol. 3, No. 2, pp. 90-109, April, 2022.

[9] K. S. Hsieh, C. M. Wang, Constructive image steganography using example-based weighted color transfer, *Journal of Information Security and Applications*, Vol. 65, pp. 1-18, March, 2022.

[10] Y. Li, Y. Jiang, Z. Li, S. T. Xia, Backdoor learning: A survey, IEEE Transactions on Neural Networks and Learning Systems, Vol. 35, No. 1, pp. 5-22, January, 2024.

[11] S. Sridhar, S. Sanagavarapu, Extending Deep Neural Categorisation Models for Recommendations by Applying Gradient Based Learning, *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*, Kuala Lumpur, Malaysia, 2021, pp. 249-254.

[12] M. A. Iqbal, Z. Wang, Z. A. Ali, S. Riaz, Automatic fish species classification using deep convolutional neural networks, *Wireless Personal Communications*, Vol. 116, No. 2, pp. 1043-1053, January, 2021.

[13] S. Mascarenhas, M. Agarwal, A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification, *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, Bengaluru, India, 2021, pp. 96-99.

[14] S. Wu, S. Zhong, Y. Liu, Deep residual learning for image steganalysis, *Multimedia tools and applications*, Vol. 77, No. 9, pp. 10437-10453, May, 2018.

[15] Q. Liu, T. Zhou, Z. Cai, Y. Yuan, M. Xu, J. Qin, M. Ma, Turning backdoors for efficient privacy protection

against image retrieval violations, *Information Processing & Management*, Vol. 60, No. 5, pp. 1-19, September, 2023.

[16] Y. Li, J. Hua, H. Wang, C. Chen, Y. Liu, DeepPayload: Black-box backdoor attack on deep learning models through neural payload injection, *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, Madrid, Spain, 2021, pp. 263-274.

[17] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278-2324, November, 1998.

[18] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *Proceedings of the Advances in Neural Information Processing Systems 25*, Lake Tahoe, Nevada, USA, 2012, pp. 1-9.

[19] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *Proceedings of the International Conference on Learning Representations*, San Diego, CA, USA, 2015, pp. 1-14.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770-778.

[21] The MITRE Corporation, MITRE ATLAS, *Backdoor ML Model*, Retrieved from https://atlas.mitre.org/techniques/AML.T0018 (last visited on 2023/10/30).

[22] P. Y. Chen, S. Liu, Holistic Adversarial Robustness of Deep Learning Models, *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, 2023, pp. 15411-15420.

[23] T. Gu, B. Dolan-Gavitt, S. Garg, *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain*, pp. 1-13, March, 2019. https://arxiv.org/abs/1708.06733

[24] E. Wenger, J. Passananti, A. N. Bhagoji, Y. Yao, H. Zheng, B. Y. Zhao, Backdoor Attacks Against Deep Learning Systems in the Physical World, P*roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 6206-6215.

[25] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks, *Proceedings of the International Conference on Neural Information Processing Systems*, Montreal, Canada, 2018, pp. 6106-6116.

[26] Y. Yao, H. Li, H. Zheng, B. Y. Zhao, Latent Backdoor Attacks on Deep Neural Networks, *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, London, United Kingdom, 2019, pp. 2041-2055.

[27] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, B. Srivastava, Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, *Proceedings of the Workshop on Artificial Intelligence Safety co-located with the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019, pp. 1-10.

[28] R. Costales, C. Mao, R. Norwitz, B. Kim, J. Yang, Live Trojan Attacks on Deep Neural Networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle, WA, USA, 2020, pp. 3460-3469.

[29] J. Lin, L. Xu, Y. Liu, X. Zhang, Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features, *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, USA, 2020, pp. 113-131.

[30] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, D. Miller, Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation, *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, New Orleans, LA, USA, 2020, pp. 97-108.

[31] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, V. Shmatikov, How To Backdoor Federated Learning, *Proceedings of Machine Learning Research*, Vol. 108, pp. 2938-2948, 2020.

[32] E. Quiring, K. Rieck, Backdooring and Poisoning Neural Networks with Image-Scaling Attacks, *Proceedings of the IEEE Security and Privacy Workshops*, San Francisco, CA, USA, 2020, pp. 41-47.

[33] R. Pang, H. Shen, X. Zhang, S. Ji, Y. Vorobeychik, X. Luo, A. Liu, T. Wang, A Tale of Evil Twins: Adversarial Inputs versus Poisoned Models, *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, Virtual Event, USA, 2020, pp. 85-99.

[34] A. S. Rakin, Z. He, D. Fan, TBT: Targeted Neural Network Attack With Bit Trojan, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 13195-13204.

[35] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, S. Xia, Rethinking the Trigger of Backdoor Attack, *Proceedings of the International Conference on Learning Representations (Withdrawn Submission)*, Virtual Event, Austria, 2021, pp. 16472-16481.

[36] E. Bagdasaryan, V. Shmatikov, Blind Backdoors in Deep Learning Models, *USENIX Security Symposium*, Virtual event, 2021, pp. 1505-1521.

[37] G. Severi, J. Meyer, S. Coull, A. Oprea, Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers, *USENIX Security Symposium*, Virtual event, 2021, pp. 1487-1504.

[38] Y. Li, Y. Li, B. Wu, L. Li, R. He, S. Lyu, Invisible Backdoor Attack with Sample-Specific Triggers, *Proceedings of the IEEE International Conference on Computer Vision*, Montreal, QC, Canada, 2021, pp. 16463-16472.

[39] A. Salem, R. Wen, M. Backes, S. Ma, Y. Zhang, Dynamic Backdoor Attacks Against Machine Learning Models, *IEEE European Symposium on Security and Privacy*, Genoa, Italy, 2022, pp. 703-718.

[40] Y. Tian, F. Suya, F. Xu, D. Evans, Stealthy Backdoors as Compression Artifacts, *IEEE Transactions on Information Forensics and Security*, Vol. 17, pp. 1372-1387, March, 2022.

[41] A. J. Zargar, Digital image watermarking using LSB technique, *International Journal of Science & Engineering Research*, Vol. 5, No. 7, pp. 202-205, July, 2014.
[42] K. S. do Prado, *Steganography: Hiding an image inside another*, Retrieved from https://towardsdatascience.com/steganography-hiding-an-image-inside-another-77ca66b2acb1 (last visited on 2023/10/20)
[43] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, The German traffic sign recognition benchmark: a multi-class classification competition, *The 2011 international joint conference on neural networks*, San Jose, CA, USA, 2011, pp. 1453-1460
[44] J. Stallkamp, M. Schlipsing, J. Salmen, C. Igel, Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition, *Neural networks*, Vol. 32, pp. 323-332, August, 2012.

## Biographies

**Candra Ahmadi** He is currently working towards the Ph.D. degree at the National Taiwan University of Science and Technology (NTUST), Taiwan. His current research interests include the Internet of Things (IoT), 5G technologies, Machine Learning, Cyber Security, and Artificial Intelligence.

**Jiann-Liang Chen**, Prof Chen joined the Department of Electrical Engineering at National Taiwan University of Science and Technology in 2009, where Prof Chen now serves as a Distinguished Professor and Dean. His research focuses on cellular mobility management, cybersecurity, personal communication systems, the Internet of Things, and AI applications.

**Yu-Ting Lin** is a student at the National Taiwan University of Science and Technology. He received a bachelor's degree in Electrical Engineering from the National Taiwan University of Science and Technology. His research interest includes artificial intelligence and information security.