CHAPTER 1
INTRODUCTION

1.1 Background

The security of information systems, particularly datasets frequently used as training data in machine learning processes, must always be maintained. Enhancing dataset security can be achieved by designing an architecture implemented within a system. This architecture includes embedding watermarks into training data (datasets) and detecting unauthorized data injections. The creation of datasets typically requires significant time and resources, making them valuable assets that need protection. Attackers can insert unauthorized data into the dataset, a phenomenon known as a backdoor attack. Such attacks can be detrimental as they affect prediction accuracy and decision-making processes. Backdoor attacks can occur at any time, making it essential for dataset owners to consistently and regularly inspect their datasets.

A demonstration of backdoor attacks [4] on facial recognition systems has shown how such attacks can lead to incorrect predictions. Previous research has explored solutions to mitigate this issue, including watermark embedding [4]. Watermarking techniques have been studied in various previous works, such as the study conducted by Wang et al. [1], which applied wet-paper coding based on deep neural networks. Their research outperformed prior studies [5–11] in terms of fidelity, robustness, and particularly capacity.

DeepSigns [6] utilizes a series of random binary strings designed so that each bit is independently and identically distributed. DeepSigns produces a deep neural network pre-trained with an embedded watermark in specific layers and generates a corresponding watermark key. This allows model owners to securely distribute their pre-trained deep neural networks with embedded watermarks. However, Wang et al. [1] pointed out that vulnerabilities still exist, particularly in training data models that lack regular manual inspection, allowing attackers to insert unauthorized data and disrupt prediction accuracy.

Cohen et al. [12] conducted research on detecting hidden message locations in images. The images used were in Joint Photographic Experts Group (JPEG) format, as this format is widely used as a medium in steganography. Cohen et al. [12] proposed ASSAF, a new neural network architecture that combines a denoising con-

volutional autoencoder with a Siamese Neural Network designed for steganography detection. The use of a Siamese Neural Network requires two identical components to process and compare images. The training dataset is crucial for distinguishing between watermarked and non-watermarked images.

Wang et al. [13] employed deep neural networks to detect hidden messages in images by leveraging strong learning and classification capabilities. Their method surpassed traditional steganalysis techniques that rely on manual feature extraction. However, deep learning-based steganalysis methods still face challenges in detection accuracy, particularly for images of arbitrary sizes and multi-source datasets. Detection efficiency is also affected by cover mismatch. To address this, Wang et al. [13] proposed a Siamese Neural Network-based method with an inverted residual structure to extract residual features from subgraphs.

In this study, watermarking is embedded directly into the parameters of a Deep Neural Network (DNN), ensuring that each dataset instance has a unique identity. Unlike the watermarking method proposed by [1], which relies on binary bit sequences, this approach embeds textual watermarks in the form of email addresses associated with registered users. The embedding process is conducted using Wet Paper Coding (WPC) combined with the Optimal Parameter Selection Strategy (OPSS). By integrating watermarking at the model parameter level, this method prevents unauthorized modifications to the Deep Neural Network and allows dataset ownership to be verified through watermark extraction.

After embedding the watermark using Wet Paper Coding (WPC) and the Optimal Parameter Selection Strategy (OPSS), the email-based watermark is utilized to detect backdoor attacks by applying a Siamese Neural Network (SNN). This model compares the extracted watermark from the network parameters with a reference email dataset and measures the similarity between them. If the similarity is high, the dataset can be confirmed as authentic, otherwise, it may indicate unauthorized modifications or potential backdoor attacks.

1.2 Research Objective

The objective of this study is to develop a system that enhances the security of Deep Neural Networks (DNN) against backdoor attacks. This is achieved by embedding a watermark in the form of an email into each training dataset (MNIST and CIFAR-10), which will then be used for DNN processing. For the watermark embedding process, this study employs the wet-paper coding method based on deep neural networks [1], which has been proven to outperform previous research in

terms of fidelity, robustness, and, particularly, capacity.

The proposed scheme aims to prevent data manipulation by attackers by embedding authenticated information through a database list of emails, ensuring that each user has a unique word embedded in the watermark. This enhances security in the watermark embedding process. Additionally, research by Cohen et al. [12] has demonstrated that Siamese Neural Networks (SNN) are effective in detecting words embedded in watermarked images, with the ability to distinguish between original and fake images. Siamese Neural Networks are a neural network architecture designed to compare and map similarities or relationships between inputs, such as images or text. This approach is expected to enable effective detection of data within the dataset (MNIST and CIFAR-10), thereby ensuring its security and integrity.

1.3 Problem Statement

The security of Deep Neural Networks (DNN) in machine learning is highly vulnerable to backdoor attacks, where attackers can inject unauthorized data to manipulate model behavior in DNNs. One of the methods developed to protect Deep Neural Networks is watermarking using Wet Paper Coding (WPC), as introduced by Wang et al. [1]. This study demonstrated that WPC offers advantages in terms of fidelity, robustness, and embedding capacity. However, this method still has limitations in watermark extraction, where the embedded text often undergoes changes due to variations in the embedding locations determined by the Optimal Parameter Selection Strategy (OPSS). As a result, the extracted watermark is often inaccurate, reducing the reliability of this method in verifying data authenticity. Therefore, based on the research by [12], this study proposes the use of a Siamese Neural Network (SNN) to detect backdoor attacks by verifying whether a dataset originates from an authorized user based on words embedded as a watermark.

Unlike previous research, this study uses an email as the embedded watermark instead of binary bits (01). The use of emails aims to enhance the uniqueness of the watermark and ensure that each training data instance can be traced back to its legitimate owner. However, the main challenge of this approach is the significant differences between the original email text and the extracted result, which can lead to errors in backdoor attack detection.

Based on these issues, this study aims to develop a more effective watermark embedding and extraction method and optimize the SNN model to improve the accuracy of backdoor attack detection based on embedded email watermarks. This

approach is expected to reduce errors in watermark extraction and enhance the system's ability to identify unauthorized modifications in training datasets.

1.4 Hypothesis

The watermark embedded in the model parameters using the WPC and OPSS methods may change during the extraction process due to factors such as variations in the words within the watermark. One of the main challenges in watermarking methods is that the extracted watermark may not be perfectly identical to the original one. Sometimes, words or symbols hidden within the watermark may undergo slight modifications due to the embedding and extraction processes. Therefore, a method is needed that not only detects the presence of a watermark but also determines whether the extracted watermark remains valid or has been tampered with.

In this regard, this study will use a Siamese Neural Network (SNN) to distinguish between valid and invalid watermarks based on the feature distance between the original and extracted watermarks. By implementing the SNN model, the system can determine whether a watermarked image remains trustworthy or has undergone modifications that affect its authenticity. This means that the SNN model will be trained to recognize subtle variations in the watermark and classify extracted results as either valid or invalid. Consequently, the model will be more resistant to noise and various types of attacks, thereby improving its accuracy in detecting manipulated watermarks.

1.5 Research Methodology

This methodology aims to develop and evaluate a system that integrates Wet Paper Coding with the Optimal Parameter Selection Strategy (OPSS) to ensure the security of Deep Neural Networks (DNNs). The methodology consists of four main stages: system design, dataset preparation (MNIST, CIFAR-10, and a registered email list), implementation, and experimental evaluation. Each stage is structured to achieve the research objectives and validate the formulated hypotheses.

The embedding module inserts a watermark in the form of an email using the Wet Paper Coding technique, ensuring that the watermark remains imperceptible by selecting only the most important model layers through OPSS calculations. The embedded watermark, which consists of an email address, enhances resistance against manipulation. The authentication system, based on a Siamese Neural Network (SNN), verifies whether the watermark originates from an authorized or unautho-

rized user. The datasets used in this study are MNIST and CIFAR-10.

The proposed system is evaluated through various tests. For the watermarking system, robustness is tested by performing fine-tuning and retraining the watermarked dataset (registered email list) using a Deep Neural Network (DNN). Meanwhile, in the Siamese Neural Network (SNN) system, evaluation is conducted using accuracy, True Positive Rate (TPR), and False Positive Rate (FPR).

All experiments are conducted on an Intel Core i7 processor with 16 GB of RAM and SSD storage. The programming implementation is carried out using Python 3.12 with the PyTorch framework. This research aims to develop a robust, efficient, and reliable watermarking system that can be applied to various use cases.

1.6 Problem Limitation

To limit the scope of this research, several problem restrictions have been identified as follows:

1. Using Google Colab Pro for watermark embedding and Windows 11 for detecting backdoor attacks using a Siamese Neural Network.
2. Implementing the system using Python 3.12.
3. Embedding an email as the watermark during the watermarking process.
4. Using the MNIST and CIFAR-10 datasets.
5. Calculating the watermark extraction results using Detection Rate and Bit Error Rate.
6. Evaluating fidelity by retraining the watermarked dataset to assess its accuracy.
7. The parameter variations for sample selection in the SNN testing are derived from the best parameter configurations during watermark embedding, including dataset (MNIST and CIFAR - 10) type, layer ratio, probability alpha, and dry block ratio, based on DR and BER results.
8. Using a registered email list containing 9 emails.