DETECTING MANIPULATED AND ADVERSARIAL IMAGES: A COMPREHENSIVE STUDY OF REAL-WORLD APPLICATIONS

by

MOHAMMED ALKHOWAITER
M.S. The University of Tulsa, United States, 2018
B.S. Prince Sattam Bin Abdulaziz University , Saudi Arabia, 2013

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Engineering
in the College of Engineering and Computer Science
at the University of Central Florida,
Orlando, Florida

Fall Term
2023

Major Professor: Cliff Zou

ABSTRACT

The great advance of communication technology comes with a rapid increase of disinformation in many kinds and shapes; manipulated images are one of the primary examples of disinformation that can affect many users. Such activity can severely impact public behavior, attitude, and belief or sway the viewers' perception in any malicious or benign direction. Additionally, adversarial attacks targeting deep learning models pose a severe risk to computer vision applications. This dissertation explores ways of detecting and resisting manipulated or adversarial attack images. The first contribution evaluates perceptual hashing (pHash) algorithms for detecting image manipulation on social media platforms like Facebook and Twitter. The study demonstrates the differences in image processing between the two platforms and proposes a new approach to find the optimal detection threshold for each algorithm. The next contribution develops a new pHash authentication to detect fake imagery on social media networks, using a self-supervised learning framework and contrastive loss. In addition, a fake image sample generator is developed to cover three major image manipulating operations (copy-move, splicing, removal). The proposed authentication technique outperforms the state-of-the-art pHash methods. The third contribution addresses the challenges of adversarial attacks to deep learning models. A new adversarial-aware deep learning system is proposed using a classical machine learning model as the secondary verification system to complement the primary deep learning model in image classification. The proposed approach outperforms current state-of-the-art adversarial defense systems. Finally, the fourth contribution fuses big data from Extra-Military resources to support military decision-making. The study proposes a workflow, reviews data availability, security, privacy, and integrity challenges, and suggests solutions. A demonstration of the proposed image authentication is introduced to prevent wrong decisions and increase integrity. Overall, the dissertation provides practical solutions for detecting manipulated and adversarial attack images and integrates our proposed solutions in supporting military decision-making workflow.

To my family.

ACKNOWLEDGMENTS

Throughout my educational journey at one of the world's top schools, UCF, I regret not taking the time to fully appreciate the school's atmosphere due to my intense focus on academic work. However, I gained significant knowledge in my field and acquired high-level research skills that will benefit me in my future endeavors. I could not have accomplished this without the help of the people around me who provided support and guidance along the way.

I am deeply grateful to my advisor, Professor Cliff Zou, who has imparted unlimited knowledge and professional and life experience since I met him at one of the UCF expos. His commitment and dedication have taught me who I should be in my career. I also want to thank the committee members who carefully examined my dissertation and provided invaluable feedback that helped me succeed.

I express my heartfelt gratitude to my parents, father Abdulaziz and mother Hessah. They have been merciful, caring, and supportive, and have sacrificed a lot to help me achieve my goals. Their unwavering belief in my abilities and constant encouragement has been the foundation upon which I have built my aspirations. They have been my silent yet most potent companions throughout this journey.

My wife Ghaida is my sanctuary in life. She has been patient and supportive, caring for our children and me during the hard days. Your love has been my refuge, your resilience my inspiration. You have stood by me, celebrating every small triumph and holding me up through every stumble.

To my sweet and wonderful daughter Jumanah, my heart's delight, thank you for being my shining star and bringing joy to our family. I am so blessed by your beautiful smile, happiness, and willingness to share your toys and games with me and others (sharing is caring). You are a shining example of love and kindness, and you are so bright that you have interacted with me and owned my heart since your first couple of months.

To my newborn son, Abdulaziz, who arrived at the end of my journey, I apologize for not being

able to show you the beautiful places where you were born. Don't worry, your older sister will tell you many stories about them. Thank you for bringing more love to our home. I hope to see you proud of your dad's achievements and wish you to make your own and even better achievements.

Lastly, I would like to acknowledge all my friends and well-wishers who have supported me with their assistance, encouragement, and kindness throughout this journey. You have been the shining stars in my scholarly journey, guiding me toward this moment of triumph and fulfillment. I dedicate this dissertation to you all with immense gratitude and love. Thank you for being a part of my life's mosaic of support and love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background

Social media platforms—Facebook, Twitter, etc.—speed up the spread of information throughout the well-connected cyber world. However, at the same time, these platforms help to forge misinformation that can quickly reach a massive number of people. Propagation of fake information and news can lead to deception and emotional distress and influence public opinions and actions. An investigation into the truth of news on Twitter from 2006 to 2017 showed that falsehood diffuses faster and more profoundly than truth [103]. The risk of misinformation increases during significant events. A study on fake images on Twitter during Hurricane Sandy (2012) [34] showed that around 90 percent of retweets were from tweets of fake images. These fake images not only mislead users, but also can contain malicious URLs.

In addition, a study [106] shows that fake images on social networks increase user engagement, regardless of the depth of manipulation. Fig. 1.1 shows a real-world example of fake image spreading over the media. The altered photograph, investigated by `Snopes.com` [28], shows a billboard to hire 'crypto bros' advertised by McDonald's. This faked image spread along with the crush of some cryptocurrencies, leaving a cruel emotion to those investors who lost their money or affected the crypto market by losing more revenue. Another manipulated image spread over the Internet shows George W. Bush at a book reading at school in Houston in 2002 holding the book upside down with a false caption [63]. The impact of manipulated images is countless on different timeline events where the damage differs based on the targets.

Most social media platform users are unaware of the risk of re-sharing information from unknown sources. Therefore, it is impossible to prevent the spread of disinformation without an automation technique, and a solution that reduces misinformation on the Internet is urgently needed. Recently, many platforms activated fake detection features to reduce or eliminate false information. For example, during the 2020 presidential election in the United States and COVID-19, social plat-

(a) (b)

(c) (d)

Figure 1.1: Originals images (a and c) and their altered copies (b and d) that spread over the network.

Source: McDonald's billboard images: `https://www.snopes.com/fact-check/crypto-mcdonalds-billboard/` President Bush images: `https://www.snopes.com/fact-check/bush-upside-book/`

forms [62, 83] labeled misinformation posts. However, there are still many technical challenges to research and conquer to win the war against disinformation spreading, especially for images.

An example of the challenges is that social media platforms automatically alter images upon sharing for many reasons, i.e., images are re-scaled and compressed to save room on the servers, and each platform shares its preferences of image sizes [68]. This means that when sharing an image, the social network will resize it to fit its preferred dimension. For instance, an image of 3000x2000 pixels will be downscaled by Facebook to 1875x1250, Instagram to 1080x720, and Twitter to 680x453. This scaling diversity is one of the image attacks distinguished systems suffer

from while authenticating the images.

Another challenge associated with image authentication systems, especially in the deep learning approach of machine learning, is the risk of adversarial attacks. Adversarial attacks involve image manipulation designed to deceive computer vision tasks, making the image appear correct to human perception [33]. These attacks can lead to harmful failures in sensitive computer vision-based applications, such as image authentication or autonomous vehicles misinterpreting a STOP sign as a SPEED LIMIT 65 sign. The demand for AI applications is increasing, which may increase the risk of these attacks if the technology is not secured before it is marketed. Therefore, researchers have been developing algorithms and systems to prevent adversarial attacks. Intelligent defense and security systems are essential to reduce or prevent the risks of adversarial attacks.

Deep neural network (DNN) theory, also called deep learning, accelerates the development of computer vision applications to advance the work presented in [2,4,50,79,102]. Unlike other ML approaches, DNNs can quickly learn complex patterns and representations from large and high-dimensional datasets. Therefore, according to a Stone study [92], DNN technology is expected to be used in an expanding range of real-world applications within the next decade. Examples of these applications include autonomous vehicles, security surveillance cameras, and health care. However, this technology faces serious security challenges due to two factors. One is the high dimensionality and complexity of the input data to DNN models, which means that it is difficult to catch all potential attacks, as adversarial attackers can insert small but sufficient perturbations to mislead the system. Second is the non-linearity in the decision boundaries of DNNs, resulting in unexpected and complex behaviors that are difficult to predict.

Nevertheless, the reliability of information gain more attraction in designing automated decision support making. The advent of big data has revolutionized how organizations handle, store, and analyze large volumes of data. It is, coupled with the development of more efficient hardware, has given way to the era of AI. Despite the limitations, these concepts have practical application in the military domain. For example, the United States military uses Joint All-Domain Command and

Control (JADC2) to integrate multiple data sources and sensors from various domains such as land, sea, air, space, and cyberspace to facilitate faster and more informed decision making. Another example is the Common Operational Picture (COP) concept proposed by the U.S. Department of Defense. This concept aims to provide shared situational awareness across all levels of an organization, from tactical to strategic, to enable more informed decision making based on a common understanding of the real-time operational environment. Additionally, the NATO community has discussed and tested the Data Lake concept through the NATO Core Data Framework (NCDF) as a solution for sharing reliable information at the right time/form among coalition partners across domains.

AI can process large amounts of data quickly and identify difficult or impossible patterns for humans to detect. As a result, the military can use AI to improve its experience in the field, optimize operations, automate tasks, make data-driven decisions, correlate data from different sources, and support countermeasures in case of threats and disasters. Command and Control (C2) can collect and combine information from different data sources to describe and understand the urban scenario and make accurate and context-aware decisions using the concept of data fusion. Due to the ongoing growth of big data and smart cities, many modern cities have a wide distribution of sensors that can collect beneficial data for urban military operations. Furthermore, people share information via social media platforms, such as Twitter, Facebook, Instagram, and YouTube, providing data in the form of text, images, or videos. This information can also improve the operational picture but poses challenges, such as data integrity, as there is no guarantee that the information provided is valid or not intentionally manipulated.

1.2 Motivation

Many researchers have studied image similarity measuring systems over the past years and introduced them for different purposes, such as image near-duplicates [48], search engines [109], image retrieval [86], image classifications [13, 27, 40, 80], and image authentication [55, 77, 99, 105, 113,

116, 117]. These systems are based on Perceptual Hashing (pHash) for image hash generation. pHash is tolerant to slight changes that do not affect the image content, such as compression, scaling, blurring, and rotation, where it gives a string representation as the traditional cryptographic hashing method provides for authentication. It generates a fingerprint of an image by analyzing and extracting features of the image that can be invariant under various attacks. These features, then, are taken to finalize the hash value. This value is compared with the tested image hash value to decide whether the tested image is similar, tampered with, or different. In contrast, traditional cryptography is sensitive to single-bit changes that is unreliable for image domain authentication on the Internet.

The development of previous works of image authentication relied on different benchmarks, such as CASIA [118], USC-SIPI [101], and PS-Battles [39]. The majority of the datasets used in these previous papers were manually crafted or applied significant alterations in the images, making the developed models unverifiable by real-world applications nor effective for small content-alteration. Therefore, the need of reevaluating these approaches on real-world applications is important.

In order to contribute to developing a robust authentication system, we conducted a review of the literature on pHash. We found that few works have been introduced for image authentication using Machine Learning (ML), such as [77], which shows that Convolutional Neural Network (CNN) is effective in developing a pHash system with high accuracy. However, at the evaluation step in [77], they used the same JPEG compression with quality factors of $\{1, 5, 10, 30, 50, 70, 90, 95\}$ at its training stage as content-preserving image operation. In real-world applications, the quality factor could be any number between 1:100, which might create an evaluation bias in their proposal.

In addition, the success of related work on image classification [37, 89] inspired us to investigate different CNN approaches for image authentication. Regardless of different CNN architecture designs, such as layer length, channels, and kernel size, the output of each CNN model provides the best image feature representation for classification. We found that these vector representations can be projected and exploited for image hashing with a smaller hash length to keep more space in

the memory. In short, the projection vectors of the last layer were used and converted into buckets using random projection of Locality Sensitive Hashing (LSH) [31].

Developing a novel authentication system using ML led to fights against the challenges of adversarial attacks. The new idea was inspired by analyzing communication war in the real world, as described below. Suppose a war scenario or simulation in which the Blue team uses satellite communication to operate its military. If the other side, the Red team, is somehow capable of modifying the Blue team's satellite communication without being detected, then the Blue team is misled and could lose the war eventually. In defending against such an attack in disruption of its communication, the Blue team could add a secondary radiotelegraphy system to complement its main satellite communication because radiotelegraphy, relying on a completely different mechanism, cannot be disrupted by the Red team's satellite attack methodology. Although radiotelegraphy using Morse code has a very limited bandwidth, it can transmit summary data that match the complete data transmitted via satellite communication. In this system, if the receiver discovers that the information between radiotelegraphy and satellite communication does not match, it can tell that a Red team satellite-based attack is ongoing and will not be fooled by the information.

Our proposed defense system against adversarial attacks uses the same philosophy as the war scenario described above. The deep learning image classification system is an analogy to satellite communication and can be compromised by various adversarial attacks. However, we propose the use of a traditional ML algorithm, such as RF, in analogy to radiotelegraphy as the secondary verification system. Although it is less accurate than a normal deep-learning image classification system, it is immune to most known adversarial attacks because it does not rely on a neural network structure. In this way, we can detect adversarial attacks easily when there is a mismatch between the outputs of the primary deep learning module and the secondary RF module.

Adapting our developed and secured authentication system to a real-world military application is an intriguing case. It requires the use of big data to support military decisions and addressing the challenges that come with it. This text aims to cover aspects often overlooked in the field and pre-

sented in a way that is easy to comprehend. We investigate the challenges, address techniques, and introduce a paradigm map to support quick and easy decision-making automation with a minimum number of mistakes.

1.3 Contributions

Our research contributions are mainly based on four phases. First, we create a new benchmark based on real-world applications. To do this, we collect images from various social media platforms [5]. Our research has shown weaknesses in image authentication algorithms, which presents an opportunity for further development. We evaluated six algorithms [78, 96, 97, 100, 104, 114] by implementing them and passing a set of images through these systems. We also introduced a metric algorithm to measure the robustness of their authentication. The testbed used in this evaluation consists of real images of the platforms that we generated. Our assessment will contribute to the development of an effective image authentication platform for images from social media in future work.

Next, inspired by [55,77] that uses ML for image authentication, we integrated a well-known CNN network [84] to enhance image hashing generation. We propose a model to detect manipulations on user-generated content and evaluate it using Facebook and Twitter images collected at the evaluation phase as a real-world application. The new system's results significantly improve compared to the existing approaches.

Third, based on a large-scale experiment and investigation, we find a ground similarity between various adversarial attacks on different deep learning models, which motivated us to develop this research work, as illustrated in Section 5.1. The main contribution of this work is the integration of the primary deep learning model with an additional traditional ML model that is not based on the neural network architecture (presented in Section 5.2). Additionally, a new defense metric for selecting the highest Top_k predicted class probabilities of an input sample is introduced in Section 5.2. The misclassification issue of DNN models is also addressed in the same section, and

an overall DNN model with improved accuracy is discussed.

Our method surpasses all other state-of-the-art defense methods in detecting multiple adversarial attacks using the CIFAR-100 dataset [49], shown in Section 5.3. A thorough discussion of our research is presented in Section 5.4, covering the solutions, the challenges, and the limitations encountered.

Finally, In term of demonstrating our secure approach of image authentication, we start defining the concept of Military Data Space (MDS), which encompasses Intra-Military Data (IMD) and Extra-Military Data (EMD). Then, we showcase the benefits of big data through use cases that focus on data fusion and image integrity mechanisms to support the military. Finally, we discuss the challenges and opportunities of using big data, emphasizing three main aspects: data fusion, security/privacy and integrity, and AI. These aspects must be taken into account to support strategic military decisions.

1.4 Organization

To this end, this dissertation makes four different contributions in four different thrusts:

Thrust 1. *Evaluating Perceptual Hashing Algorithms in Detecting Image Manipulation Over Social Media Platforms*: In the first thrust, a real-world dataset is collected from Facebook and Twitter social media platforms and proposed under the name named (SMPI[1]) in order to make a new evaluation of the state-of-the-art existing image authentication methods that use pHash concept. We demonstrate the results on different metrics, including a new method for selecting the best optimal detection threshold based on the SMPI dataset.

Thrust 2. *Image Authentication Using Self-Supervised Learning To Detect Manipulation Over Social Network Platforms*: In this thrust, we introduce a novel image authentication system using A self-supervised framework machine learning. First, an alteration technique is de-

[1] https://github.com/mohammedkw11/SMPI

signed to better detect copy-move, splicing, and removal operations for the pre-processing phase. Then, we employ a CNN model to construct and train the dataset to obtain the images' characteristics. After that, a Locality Sensitive Hashing (LSH) is integrated with a deep CNN at the test phase to construct the final hash. Our method is compared with state-of-the-art methods using SMPI dataset [5], IMD2020 [67], and COVERAGE [110] and showed the best performance among them all for detecting manipulated images.

Thrust 3. *Adversarial Attacks Defense Using Secondary System Detection*: In this thrust, we fill the security threats with the CNN model that we use in the proposed image authentication system. We study the adversarial attack to provide a defense technique. We first analyze the differences between the attacks, analyze different CNN models on different attacks, and then investigate the solution based on the findings inspired by a classical technique used during the beginning of satellite communications. A new defense metric is introduced to select the highest class probabilities predicted by Top_k of an input sample, depending on the application security preferences.

Thrust 4. *The Use of Perceptual Hashing Algorithms in Real-World Applications: Military Case Scenario*: In thrust, we dive into the use of big data to facilitate decision-making in military space operations. We propose a comprehensive approach that covers every aspect of the data life-cycle, starting from data collection, transformation, validation, and security up to its finalization for immediate use in decision-making. Our approach is integrated with the proposed image authentication model to reduce misinformation that could potentially influence the decision-making framework. Our main objective is to highlight the importance of our authentication ideas by demonstrating their effectiveness in a real-world example, using the military domain as a case study.

The dissertation is structured as follows: We start by reviewing the relevant literature and highlighting notable related works in chapter 2. In chapter 3, we present an extensive analysis of six pHash algorithms for image authentication. Chapter 4 covers our proposed pHash model, while

in Chapter 5, we introduce our idea for detecting adversarial attacks. Lastly, we demonstrate the real-world application of our image authentication approach in chapter 6.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we discuss the related literature, starting with work related to image authentication approaches. Followed by introducing a new model of authentication based on self-supervised technique. Then, we propose our adversarial defense approach. Finally, we show our scenario of authentication case for military applications.

2.1 Evaluating Image authentication Methods

Many academic researchers [97], [114] work on image authentication using a perceptual hashing approach and reached a high value of robustness in preventing one or more image attacks. In the survey paper [26], Du et al. classified the images attacks into two branches. First, content-preserving manipulations that do not change an image's content, such as compression, brightness reduction, and scaling. Second, content-changing manipulations that change an image's content i.e., removing image objects (persons, objects, etc.), moving image elements, changing their positions, adding new objects, etc.

Under content-preserving manipulations path, this paper [26] classified the techniques of image hashing based on five approaches proposed in the publications: (1) Invariant feature transform methods that extract image features from transform domains and then make use of the coefficients to create the hash; (2) Local feature points such as corners, edges, salient regions, etc.; (3) Dimension reduction method where robust features are extracted from embedding the low-level features of the high dimensional space into a lower dimension, (4) Statistical feature-based approach where calculation of image statistics, such as histogram and mean, are the feature from the image; and (5) Learning-based method that applies machine learning for image feature extracting and authentication.

Based on reviewing perceptual hash algorithms and their papers, any perceptual hash algorithm can be represented by three stages in image authentication as illustrated in Fig. 2.1, which will be

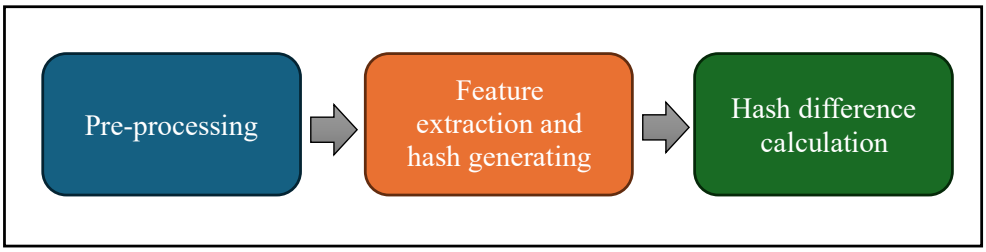


Figure 2.1: Stages of perceptual hash image authentication.

explained in detail below.

Preparing the images: (pre-processing) stage. All six approaches pass images through different enhancement lines to normalize the images in the best representation and for robust features. The common enhancement is resizing the image into fixed and small sizes to speed up the operation. Usually, it is resized into square $M \times M$, i.e., 128×128 or 224×224 pixels using the bilinear interpolation method. Another enhancement is converting the color space domain from one form to another, such as from RGB to CIELAB ($L * a * b*$) [100], and $L*$ is only used for feature extraction because it matches the human perception of lightness and is more stable. The grey-scale conversion is desirable for most developments due to its simplicity when dealing with $1D$ instead of $3D$ channels in RGB. Filters sometimes are applied, such as Gaussian and bilateral filters [114] to remove regular noises.

Feature extraction and encoding. Choosing among the best state-of-art perceptual hash algorithm modules [78, 96, 97, 100, 104, 114], we focus on reviewing and re-implementing these six modules that cover state-of-art algorithms DCT, Wavelet, RPIVD, QFT, SimCLR. Each algorithm is adaptive in an image hashing scheme design to generate the hash value that will be used at the next stage. Their authors select the size of the hash in each algorithm as optimal representation, and respectfully short definitions are provided below.

1) DCT: Discrete Cosine Transform (DCT) [97] is one of the popular algorithms that was well implemented for image compression and hashing in the last two decades and used by [75, 76, 98]. This method is invariant feature transform-based, i.e., it can represent the image in uniqueness with

small data. Basically, DCT operates on a function at a finite number of discrete data points. These data were evaluated in terms of the sum of cosine functions with different frequencies to convert it from the spatial domain to the frequency domain. The hash size is eight vectors of byte-sized integers or 64 bits.

2) Wavelet: the introduction of the DCT led to the development of wavelet coding DWT [78], which also takes a large volume of research. The wavelet is also a frequency-based technique but uses temporal details to overcome the drawbacks of DCT. The hash size is eight vectors of byte-sized integers or 64 bits.

3) Visual Model-Based: In this paper [104], Wang proposed a perceptual image hash method for content authentication. They combine a statistical feature-based approach with visual perception using Watson's visual model theory. Watson's visual model is used in order to preserve sensitive features that are important for humans perceiving image content processing. On the other hand, key-point-based features and image-block-based features are used to generate the intermediate hash by extracting key-point-based features using the input image to SIFT algorithm. To achieve this, the proposed method comprises two main stages: 1) Hash Generation Algorithm and 2) Tampering Detection and Tampering Localization. The module is against different image attacks, including geometric attacks. The hash size is 50 vectors of floating or 1600 bits.

4) RPIVD: Tang [100] designed a robust image hashing using Ring Partition and Invariant Vector Distance (RPIVD) that is considered a statistical feature-based approach. Their module is processed in three stages: (1) preparing the image by bilinear interpolation resizing into $M \times M$, low pass filtering, and converting the color space to CIE $L * a * b*$ in order to take $L*$, (2) partitioning the image into equal rings which they choose 512 rings, (3) applying four statistical measures (mean, variance, skewness, and kurtosis) to each ring. This paper provides robust image hashing against several attacks but most strongly against rotation. The hash size selected for this evaluation is the optimal representation from [16], which is 40 vectors with 11 bits representation of each vector or 440 bits in total.

5) QFT: Yan in the work [114] introduced another perceptual hashing approach that used Quaternion Fourier Transform (QFT) to construct feature hash, and QFMT to construct geometric hash. QFT is considered an invariant feature transform-based method that extracts image features from color and structural information to form a quaternion image to produce the hash. QFT can defend against common image attacks and performs well at detecting and locating various types of attacks. The hash size is 40 vectors of floating or 1280 bits.

6) SimCLR: Chen et al. [96] proposed a self-supervised learning framework using contrastive learning to learn better visual representation. The model consists of three main components. The model uses data augmentation as a critical part for the model to map similar images to proper embedding representation. To improve the learned representation, they introduced fully connected non-linear layers between the representation and contrastive loss. Moreover, large batches and more training steps are essential parts. Moreover, SimCLR uses contrastive loss in which the model tries to maximize positive examples, which are two images augmented from the same image and share similar contents. At the same time, contrastive loss minimizes negative examples, which are the reset of the images in the batch to prevent the model from mapping all the images to a constant value. Therefore, SimCLR naturally fits as a prominent model in image authentication tasks.

Similarity metric. After the generation of perceptual hashing, images can be authenticated based on the different values. There are multiple metrics for perceptual hashing comparison; however, most of the approaches as mentioned earlier follow one of two metrics for measuring: Hamming distance [78, 97] and Euclidean distance [100, 104, 114]. We integrated a Locality Sensitivity Hashing (LSH) [69] and Hamming distance in [96] to work on Neural Hash framework as in [45]. Further details will be provided in the IV section. The threshold T is set by each algorithm to distinguish whether the distance value is less or equal to the threshold value in order to decide whether the images are similar, tampered with, or different. Smaller T is better and more secure, especially for image authentication and collision probability reduction. Therefore, the best system algorithms can authenticate the received image with a small T and give the tampered images a high

distance value.

This work.. We explore a hypothesis of whether perceptual hashing algorithms can effectively authenticate images on popular social media platforms, despite the normal image processing conducted by these social media platforms on user-uploaded images. We create a dataset and evaluate state-of-art perceptual hashing algorithms over two popular social media platforms: Facebook and Twitter. We choose these two platforms for study because other platforms put tight restrictions on automatic image uploading/downloading, which prevents us from conducting large-scale image testing and experiments. For example, Instagram needs to upload and download the individual image manually, while Facebook, on the other hand, allows us to share a group of 100 images at one time. Twitter API allows developers to post 100 images per hour automatically.

2.2 Proposed Image Authentication Approach Using Machine Learning

Many image authentication works are shallow approaches that use traditional engineered algorithms. However, in recent years awareness of machine learning approaches increased with the success of deep learning models such as AlexNet in ImageNet classification challenge and promising results in other computer vision problems. Therefore, it is reasonable to say that most image authentication algorithms are built on top of shallow or machine learning paradigms. The following is an overview of most recognized works on images pHash under these two approaches.

Shallow approach. The followers of this model such as Discrete Cosine Transform (DCT) [25] provides an excellent work in representing images from different scales with small digits (e.g., 64 bits) to express the number of discrete data points. These data were evaluated in terms of the sum of cosine functions with different frequencies to convert it from the spatial domain to the frequency domain. Ring Partition and Invariant Vector Distance (RPIVD) is another shallow model introduced by [99]. They divided the image into rings and applied four statistical measures (mean, variance, skewness, and kurtosis) to each ring to extract the features. Both models are effective on image authentication based on CASIA [118], USC-SIPI [101] datasets but limited at

SMPI dataset [5].

The authors in [105] provide a perceptual image hashing method by combining a statistical feature-based approach with visual perception using Watson's visual model theory. The statistical feature-based generated by extracting key-point-based features using the input image to scale-invariant feature transform (SIFT) algorithm. The visual perception is received using Watson's visual model to preserve sensitive features that are important for humans perceiving image content processing. The accuracy of this model overcame the [25, 99] on the same benchmark ground.

Machine Learning Approach. Learned algorithms on the other side are trending these years on image classification, retrieval, and authentication since this approach extracts better feature vectors. Reference [55] proposes a data-driven image fingerprinting algorithm based on a neural network approach with two training stages: pre-trained and fine-tuning. The first stage uses a Denoising Autoencoder (DAE) to restore a distorted image to its original state. There are 72 distorted images for each original image, including nine different operations, such as JPEG compression and Gaussian noise. Each function has different strength parameters that generated the 72 copies. The main goal of this network is to reduce the discrepancies between original and distorted images. The fine-tuning approach further reduces the overlap between the probability density curves of fingerprint distances calculated from perceptually identical and irrelevant image pairings. This method is akin to restricting fingerprints into a region similar to the original image for distorted images.

The authors of [77] introduce an image pHash scheme based on the CNN framework for feature extraction and a fully connected layer at the end of the network for final image hash sequence constructing. The CNN model contains five convolutional and five pooling layers, generating 256 feature vectors, and is reduced by the fully connected layer into 50 vectors. The proposed work added four constraints. The first two constraints are added at the feature map after processing convolutional layers, the ReLU layer, and max pooling by calculating the Mean Squared Error (MSE) of identical images with perceptually identical (distorted copies) and identical images with distinct pairs. The other two constraints went before final hash construction. All four constraints

were added onto the total cost function with weight allocation. The 3,000 samples of the dataset for training are collected from COCO [57], where each image generates 64 distorted copies, and the distinct copies are paired with random images for a total of 405,000 images.

This Work. we introduced a robust authentication system targeting image authentication on social media networks. We built an alteration generator that simulates the three real-world image alteration attacks (copy-move, splicing, and removal). Also, a new data augmentation technique was included during the training process to generate distorted copies of the original.

2.3 Adversarial Attacks and Defenses

we briefly review state-of-the-art existing works on adversarial attacks and defenses. We also study the competitive detector methods that we compare our work with. These models are DkNN [70], LID [59], Mahalanibis [54], and NNIF [19].

2.3.1 Adversarial Attacks

In the past few years, many adversarial attacks have been proposed; the most common attack proposed by [33] is called the fast gradient sign method (FGSM). This attack adds a small perturbation to the target image in the direction of the gradient of the loss function with respect to the human-perception content in order to misclassify the trained targeted model. It is a white-box attack where the attacker fully knows the deep learning model, including its architecture, parameters, and training data. Later, a more efficient attack known as Deepfool [64] finds the smallest perturbation necessary to cause a DNN to misclassify an input image, which increases the attack success rate compared to FGSM.

The potential of deceiving DNN models increased significantly over the past few years of adversarial attack development. Today, imperceptible perturbations can be added to input images with the flexibility of adjusting the attack goal to either a white box or a black box, such as the one pro-

posed in [15] and named after its founders, Carlini and Wagner (CW) attack. This attack uses an optimization algorithm in order to find the smallest perturbation that minimizes a loss function that balances the size of the perturbation with the misclassification success rate. Moreover, the attack has the ability to incorporate constraints on the perturbation, such as by limiting the magnitude of the perturbation or restricting the pixel values of the perturbed image. The power of this attack raises the challenge of defense solutions against multiple attacks at once.

The white-box approach becomes more desirable for adversaries, as it was introduced by [60] and is known as the complete white-box adversary. Researchers found that the projected gradient descent (PGD) can lift any constraints on the amount of time and effort the attacker can put into finding the best attack. The iterative feature of the PGD attack makes it more effective than other attacks, such as FGSM, in finding imperceptible adversarial examples. The variety and effectiveness of adversarial attacks open a wide range of areas for researchers to develop different attacks, such as in [16, 71], and to find defense mechanisms on the other side.

2.3.2 Adversarial Defenses

The authors of [1] categorized the adversarial defense mechanisms in computer vision into three approaches. The first approach targets the deep learning model by making modifications to the model itself in order to make it more resistant to adversarial attacks. This approach was initially employed by researchers Szegedy and Goodfellow [33, 95] in 2013 and 2014, respectively. Years later, Madry [60] delved deeper into this approach by studying the robustness of neural networks against adversarial attacks from a theoretical standpoint, using robust optimization techniques. Despite its limitations, as discussed in [60], adversarial training has garnered considerable attention from the research community. In [107], a new defense algorithm called Misclassification Aware adveRsarial Training (MART) was proposed. It distinguishes between misclassified and correctly classified examples during the training process. In another study [85], researchers suggested using dropout scheduling to enhance the efficiency of adversarial training when employing single-step

methods. The authors of [66] proposed a self-supervised adversarial training method, while the authors of [18] analyzed adversarial training for self-supervision by incorporating it into pretraining.

The second approach is a defense that targets the inputs to the model by cleaning inputs to make them benign for the target model. Ref. [46] proposed ComDefend, which consists of a compression convolutional neural network (ComCNN) and a reconstruction convolutional neural network (RecCNN). The ComCNN model compresses the input image to maintain the original image structure information and purify any added perturbation. The RecCNN model, on the other hand, reconstructs the output of ComCNN to a high quality. This approach achieved high accuracy in defending against multiple adversarial attacks. GAN architecture is another technique of input transformation introduced by [87]. Their method, Defense-GAN, learns the distribution of clean images. In other words, it generates an output image close to the input image without containing the potential adversarial perturbation.

The third approach is a defense involving the addition of external modules (mainly detectors) to the target model. Among adversarial defense/detection techniques, [70] inserted a K-nearest neighbors model (k-NN) at every layer of the pretrained DNN model to estimate better prediction, confidence, and credibility for a given test sample. Afterward, a calibration dataset was used to compute the non-conformity of every test sample for a specific label (j). This involved counting the number of nearest neighbors along the DNN layer that differed from the chosen label (j). The researchers discovered that in cases in which an adversarial attack was launched on a test sample, the true label exhibited less similarity with the k-NN labels derived from the DNN activations across the layers.

The research in [59] characterized the properties of regions named adversarial subspaces by focusing on the dimensional properties using the local intrinsic dimensionality (LID). The LID method evaluates the space-filling capability of the area around a reference by analyzing the distance between the sample and its neighboring points. A classifier was trained using a dataset comprising three types of examples: adversarial as a positive class and normal and noisy (non-adversarial) as

a negative class. The features of each sample associated with each category were then constructed using the LID score calculated at every DNN layer. Finally, a logistic regression (LR) model was fitted on the LID features for the adversarial detection task.

Researchers in [54] developed generative classifiers that could detect adversarial examples by utilizing DNN activations from every layer of the training set. They used a confidence score that relied on Mahalanobis distance. First, they found the mean and covariance of activations for each class and layer. Then, they measured the Mahalanobis distance between a test sample and its nearest class-conditional Gaussian using Gaussian distributions. These distances served as features to train a logistic regression classifier. The authors found that, compared to using the Euclidean distance employed in [59], the Mahalanobis distance was significantly more effective in detecting adversarial examples and resulted in improved detection results.

In a study by Cohen et al. [19], the authors utilized an influence function to create an external adversarial detector. This function calculates how much of each training sample affects the validation data, resulting in sample influence scores. Using these scores, they identified the most supportive training instances for the validation samples. To compute a ranking of the supportive training samples, a k-NN model is also fitted on the model activations. According to their claims, supportive samples are highly correlated with the nearest neighbors of clean test samples, whereas weak correlations were found for adversarial inputs.

This Work. We introduce a simple adversarial attacks defense model based on a classical ML model. This new model shows a robust performance comparing to existed state-of-art models. Challenges are found in this approach which we tackle by developing an adaptive technique in select a threshold of security level based on applications desires.

2.4 Case Scenario of Image Authentication in Military Space Application

Big Data is widely applied today, supporting various applications in different fields. In military scenarios, we noticed a lack of discussion regarding using big data, including collecting, process-

ing, fusion, analysis, and securing to create qualitative information to design accurate and robust systems that support strategic decisions. Here we explore the use of the big military data space from both *intra* Intra-Military Data (IMD) and *extra*-data Extra-Military Data (EMD) perspectives. Also, we address the challenges of securing these data. The idea is to identify the context of big data and its security in the military and the recent proposed solutions that benefit from it.

Intra-Military Data. Some challenges of big data in the military field are presented in the literature, such as operational security, hardening against vulnerabilities, and data reliability [21, 51, 112], and also discussed in the NATO community (IST-160, and IST-173). Incorporating autonomous isolation with little connection to the outside world EMD can limit the free flow of big data, demanding creative ways to utilize big data while maintaining the autonomy and protection of the deployed systems. In this direction, Common Operational Picture (COP) and Joint All-Domain Command and Control (JADC2) have guided researchers and industries toward using and fusing data from different military entities supporting strategic decisions.

Kun et al. [51] present a detailed technical plan to construct a big data platform for scientific research in military enterprises. The platform can establish multi-level data channels across departments and promote multi-level data management and control. It can also collect, organize, process, analyze, and secure enterprise data resources, convert data into knowledge, and offer services such as decision support, product innovation, quality control, process optimization, service support, and risk management and control.

Xu et al. [112] discuss the importance of data science to achieve information superiority in contemporary warfare. A systematic literature review focuses on the opportunities or risks of data science in military decision-making at different levels of war (i.e., strategic, operational, and tactical). The study found a relatively large focus on data science risks in social science literature, which could influence political and military policymakers. However, these risks are hardly addressed in formal scientific literature. Moreover, this study points out the lack of attention to the operational and strategic levels compared to the tactical level, creating a research gap. The absence of studies, in

our view, comes from the lack of connection between IMD and EMD that may support operational and strategic decisions.

Extra-Military Data. The concept of heterogeneous data fusion contributes to the creation of big data and is widely applied in various research areas to develop solutions that improve the quality of data. Data fusion is a technique to combine multiple records describing the same object or situation into a consistent and clean representation. The main goal of this technique is to produce complete, concise and consistent data by removing uncertain or conflicting data values from a set of given information. In general, data fusion is challenging due to the data semantic heterogeneity and the different spatio-temporal aspects [81].

In a military context, heterogeneous data fusion can be used to increase the available data and therefore supports the design of military information systems. Those systems can lead to information superiority and information awareness, important aspects, especially in a complex urban warfare situation or counter-terrorism. Military information systems require a robust design regarding accuracy, reliability and credibility, because they handle human lives, exploit heterogeneous data sources and involve limited resources [12]. Data fusion supports the design of those systems by reducing possible information overload, improving the accuracy and exploiting partial or uncertain knowledge. Multi-sensor data fusion for military applications is the subject of recent literature on network communications and information systems. These investigations discuss methods and frameworks to achieve the fusion of data acquired from multiple sensors to support military applications. Multi-Sensor Data Fusion (MSDF) can be used to improve the accuracy of target tracking in a tactical scenario, supporting military operations by improving the assessment of the situation and possible threads [58].

Moreover, Location based Social Media (LBSM) can be included within data fusion procedures, as discussed by Rettore et al. [82, 88], providing two services that can improve the description of traffic conditions and detect road incidents. They introduced a framework capable of fusing heterogeneous data to output a more descriptive and enriched transportation data, but instead solely

relying on traffic- and vehicular data, they fused information from navigation systems and social media (Twitter). Besides this spatiotemporal grouping of data, improving the description of traffic conditions (T-MAPS), they furthermore used learning-based approaches for their incident detection system (T-Incident). Although these systems contribute to a civilian use case, the concept of LBSM may be used within the civilian field, to enlarge the amount of data in certain situations, leading to information superiority.

Image Authentication. Securing gathered information is a challenge in every system. In EMD and IMD scenario, the important of protecting the information increased since the goal of the module is decision-based making system. It received high sensitive data in real-time to might create a real-time decision. In our investigation in this work, we focus on one security concept, image authentication, and highlights the challenges and the opportunities of security in general.

The goal of image authentication is to verify an image's authenticity and ensure that it has not been manipulated. Recently, tools for fabricating images have increased the challenge for trusting and verifying images, starting with *Photoshop* and ending today with the image generation by artificial intelligence tools such as the platform provided by *OpenAI*. Last decade, multiple technologies were introduced to validate images using watermarking, digital signature, or pHash [3]. In watermarking, an image will have an add-on piece of digital information visible or hidden to protect digital media content, providing a way to track and verify the ownership and authenticity of digital images. There are some limitations in using this technology, such as reducing the quality of the images by embedding a photo inside a photo. Also, watermarking might be reduced or removed by advanced tools or image processing like compression and resizing.

The digital signature-based method is another authentication preservation technique where each image has a secure, bit-sensitive, and unaltered hash of string generated using one of the popular hash mechanisms such as MD5 or Sha-2. This approach works under a Public Key Infrastructure (PKI) and a secure digital key which requires each party to exchange their keys before sending or receiving the object. The limitation in using this high-security mechanism is that images forwarded

over the internet through different media platforms often go through multiple normal image processing procedures (such as compression or resizing) that alter bit-wise data but do not change the image content, making such a strict data verification method invalid.

On the other hand, pHash is an alternative approach that can be flexible to most normal image operations, such as compression and blurring, while at the same time be sensitive to image content-changing, such as copy-move, splicing, or removal. It generates a fixed length of string for the image and can be used in social media platforms as demonstrated in paper [3].

This Work. We study the type of gathered information from military space based on the availability. After that, we design a data fusion framework of collecting the data, preparing, processing, protecting, and delivering. We study the security challenges and opportunities, and integrating our image authentication in data fusion system to demonstrate the complete framework.

CHAPTER 3: EVALUATING PERCEPTUAL HASHING ALGORITHMS IN DETECTING IMAGE MANIPULATION OVER SOCIAL MEDIA PLATFORMS

In this chapter, we evaluate the effectiveness of state-of-the-art image authentication algorithms and address their weaknesses in the following steps. Firstly, we introduce a new dataset that contains real and manipulated images from a real-world environment. We collect this dataset from two of the most famous social media platforms. Then, we present our evaluation methodology and provide the outcomes of the algorithms based on their original settings. Next, we apply our new assessment method and present the results obtained from it. Finally, we conclude with a discussion statement.[1]

3.1 Methodologies

Social network platforms, Facebook and Twitter, apply image processing and enhancements such as scaling, compression, brightness reduction, and contrast. Each platform has different image effects upon posting for storage preference and transmission. To analyze the image processing operation by each platform, we upload an image on the platform, and then download the image again. From these steps, we found that each downloaded image has different sizing, scale, smoothness, and quality for different social media platforms.

The scheme of the evaluation approach that we follow is shown in Fig. 3.1, and six perceptual hash approaches are evaluated for the performance comparison. To prepare for evaluation, the first step is to generate the dataset SMPI. Creating the SMPI dataset consists of four stages. First, we randomly select 6497 images from Holopix50k [42] dataset that was originally collected from users of the Holopix mobile social platform due to the large and variant scales of the images. Secondly,

[1]The contents of this Chapter are based on our publication to IEEE CSR 2022 [5]

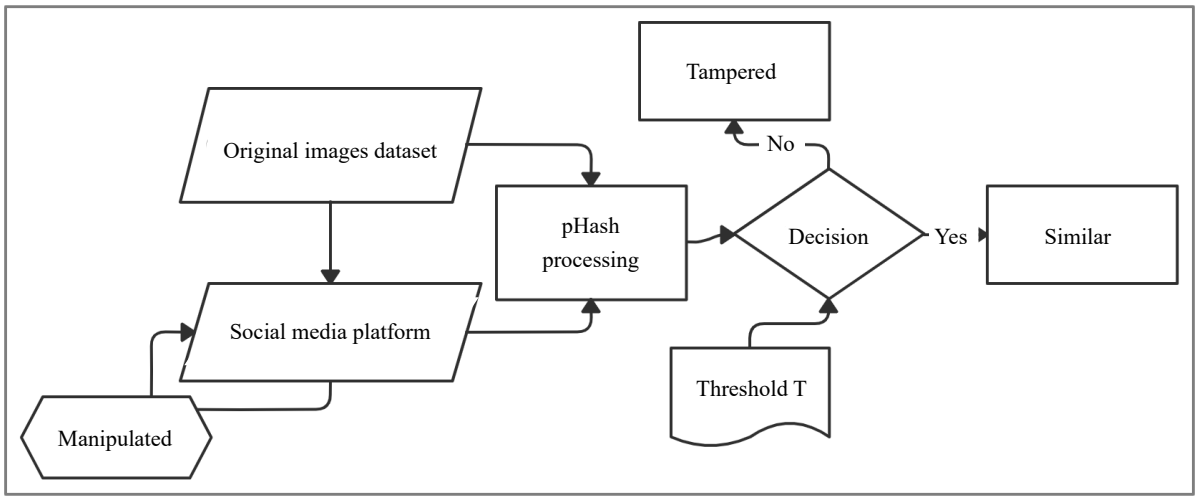


Figure 3.1: Methodology used by our proposed authentication platform for social media images.

each image is first shared on Facebook and Twitter and then downloaded. Third, the downloaded image is manipulated randomly by one of these attacks: copy-move, splicing, or removal. Finally, the manipulated image is shared again to each social media platform and then downloaded again. After this four-stage operation, we have three copies of each image: the original image (from the first step), the shared image of the original (after the second step), and the altered/shared image (after the fourth step). In total, we collect over 19K images, where each category consisting of 6497 samples.

The three image manipulation attacks mentioned above are further explained in the following. *Splicing* is designed to cut a random part of a different image and paste it randomly onto the target image. *Copy-move* copies a part of an image randomly and pasts it on a different location randomly of the same image. Finally, *removal* selects a part of the image randomly and applies a 5x5 kernel simple blurring filter 50 times on the same location. The dimension of all three types of alterations is randomly chosen where the length and width have to be larger than 4x4 pixels and smaller than the size of the entire image that receives the alteration. Figure 4 represents samples of the three categories of images in our dataset and of the three different image alterations.

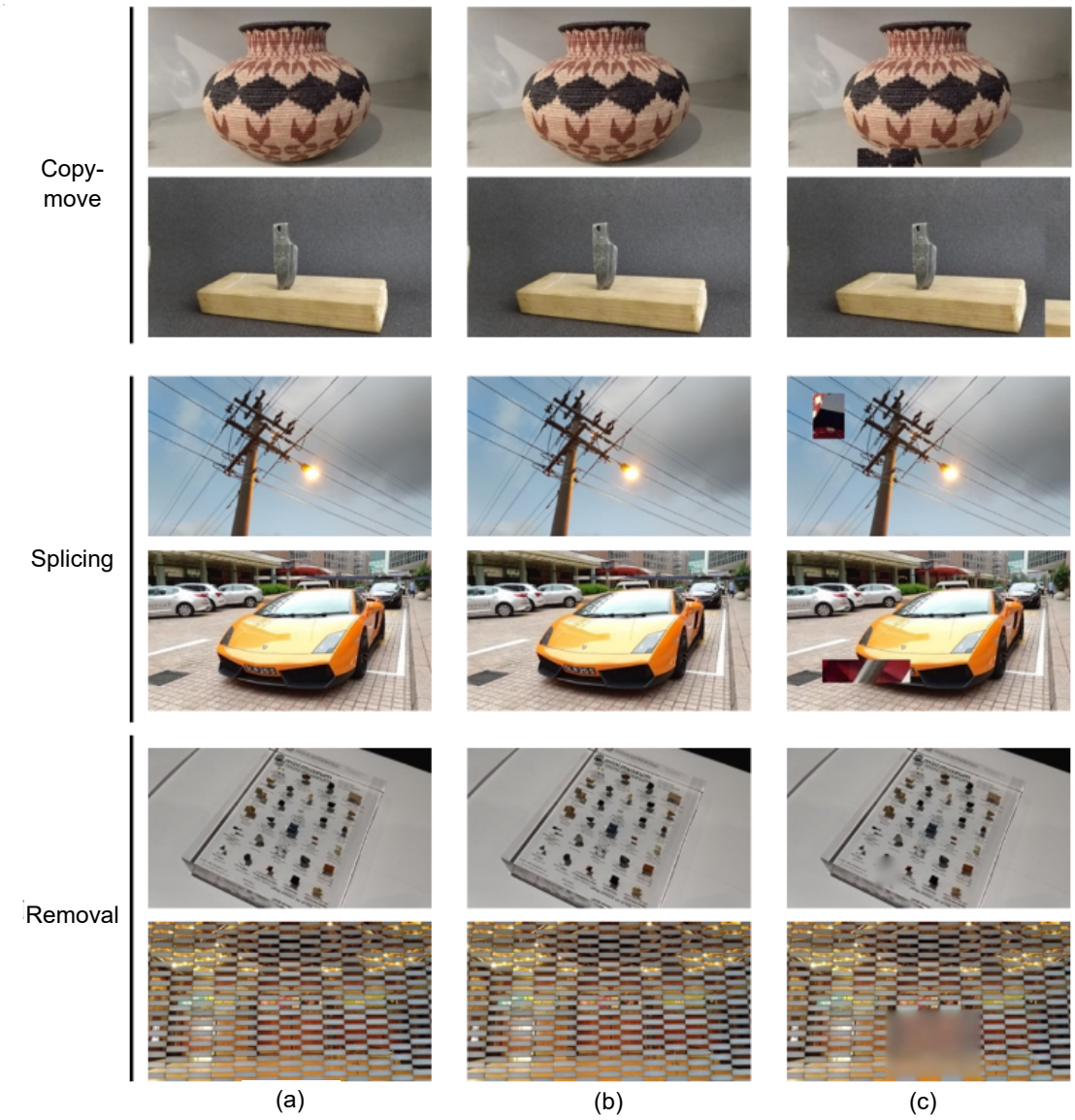


Figure 3.2: Sample of images in the dataset for evaluation: (a) original images; (b) images posted and then downloaded as is; (c) images tampered, posted, and then downloaded.

Source: Group (a) are original samples from Holopix50k Dataset.

Our assumption for the authentication system is that we have an authentication platform for social media images. The first publisher can post and generate the original perceptual hash of the image before posting it to social platforms. The platform's users could check the validation of the image by checking the image with the authentication platform by posting the image there. The

outputs for the checking will represent a message of authentic or tampered. Dissimilar images for this evaluation are ignored because all the authentication algorithms are simply successful in the detection.

The second step of the approach is preparing the algorithms to generate the perceptual hash for images. We choose the six modules based on state-of-art approaches [25, 25, 78, 96, 100, 114], the most recognized algorithms for the last decade and the technique that the algorithms follow among five approaches that [26] classified. The majority of approaches are reverse-engineered [25, 96, 100, 114] and [25, 78] are provided as open sources [14, 115]. The decision-making is chosen based on the thresholds set by the authors of [25,25,78,100,114] which are 0.25, 0.004, 5.0, 200, and 0.8, respectively, except SimCLR model that has never been used in image authentication tasks. Therefore, we selected T = 0.04 based on the best result we obtained, as it is described in part B of the IV section.

Afterward, the evaluation of the proposed methods was measured by four metrics approaches false-negative rate (FNR) and false-positive rate (FPR), gap difference $diff$, accuracy, and average time processing.

1) FNR-FPR: Equation 3.1 and Equation 3.2 represent the false-negative rate (FNR) and false positive rate (FPR) from [120] respectfully. FN is calculated by dividing FN over the summation of FN and true positive (TP). In contrast, FP is calculated by dividing FP over the summation of FP and true negative (TN). The lower rate of either FN or FP indicates the better robustness of the authentication capability of a perceptual hash algorithm.

$$FNR = FN/(FN + TP) \tag{3.1}$$

$$FPR = FP/(FP + TN) \tag{3.2}$$

2) $diff$*:* the division of average altered images hashing over average similar images hashing as

shown by Equation 3.3:

$$diff = avg(ph_{alter})/avg(ph_{similar}) \tag{3.3}$$

The perceptual hash distance, $ph_{similar}$ in Equation 3.3 refers to the Hamming distance or Euclidean distance value between the original image (Fig. 3.2a) and social media downloaded and unaltered image (Fig. 3.2b). The smaller value of $ph_{similar}$ is better, which means the perceptual hash algorithm can detect that the downloaded image from the social media platform is unaltered from the original image beside the image processing that social media platforms add upon posting such as compression and resizing. On the other hand, the perceptual hash distance, ph_{alter} refers to the Hamming distance or Euclidean distance value between the original image (Fig. 3.2a) and social media downloaded and altered image (Fig. 3.2c). The larger value of ph_{alter} is better, which means the perceptual hash algorithm can detect an altered image used in the social media platform. The avg in the equation refers to the average of all images' values in similar and alteration tests.

Furthermore, the $diff$ value calculated in Equation 3.3 reflects how well the perceptual hash algorithm can detect image alteration when an image is used on a social media platform. A good perceptual hash algorithm should have a larger $diff$ value for better decision-making in detecting image alteration in social media platforms. On the other hand, a smaller $diff$ value will make it harder to choose the detection threshold, hence, increasing the false-negative and false-positive rates.

3) Accuracy: Equation 3.4 calculates the accuracy of each algorithm based on the values of TP, TN, FN, and FP.

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.4}$$

4) average processing time: the average time that each run of generating perceptual hashing takes. This metric is affected by the processor resource and the environment we use during evaluation.

The processor is a 2.7 GHz Dual-Core Intel Core i5, the memory is 8 GB 1867 MHz DDR3, and the operating system is macOS Monterey.

3.2 Experimental Results

This section uses our proposed evaluation scheme in the previous section to evaluate those seven perceptual hash algorithms. We implemented and tested the scheme using Python (3.8.5) and some open source libraries such as OpenCV [14] to build [25] and [78]. In addition, we implemented a Neural Hash scheme which consists of a neural network that extracts image features and maps them into a vector with a fixed length. Then, the vector is fed to LSH, which maps each vector to a specific bucket where similar images have similar vectors mapped to the same bucket. The last step is to convert these matrix-vector buckets to binary hash using hamming distance. We used a SimCLR pre-trained on Imagenet [23] as in [96] to map the images into a vector of 128 length. Then this vector is mapped to 1024 bits. All our implementation for SimCLR-LSH used Pytorch framework [72]. For the remaining algorithms, we re-implemented them as they are described in these publications [25, 100, 114].

Evaluation based on algorithms' original decision thresholds set by their authors. Following the scheme in Fig. 3.1, we generated 155,928 tests equally distributed by the six modules. These tests are calculated by using the original images from the SMPI dataset. Each image calculated twice: one to calculate the $ph_{similar}$ (Fig. 3.2a Vs Fig. 3.2b) and the second for ph_{alter} (Fig. 3.2a Vs Fig. 3.2c) which creates 12994 tests for each platform and in total of 25988 tests for the two platforms. Afterward, these outputs are summarized and shown at Table 3.1. The table's minimum score (min) means the lowest perceptual hashes value among the 6497 tests at each approach and platform. The average (avg) indicates the average of 6497 tests under each algorithm and platform, and finally, the maximum (max) refers to the highest perceptual hash among the 6497 evaluations at each algorithm and platform.

Table 3.1 shows the summary of the perceptual hashing distance of these six algorithm modules in

Table 3.1: pHash similarity score between the original images (Fig. 3.2a), the posted images (Fig. 3.2b), and altered images (Fig. 3.2c) on social media platforms.

Algorithm	T	**Facebook**						**Twitter**					
		min		***avg***		***max***		***min***		***avg***		***max***	
		a&b	a&c	a&b	a&c	a&b	a&c	a&b	a&c	a&b	a&c	a&b	a&c
DCT [25]	0.25	0.0	0.0	0.04	0.28	0.5	1	0.0	0.0	0.0	0.27	0.0	1
Wavelet [78]	0.004	0.0	0.0	0.0004	0.02	0.03	0.5	0.0	0.0	0.0	0.025	0.0	0.68
Vsul M-B [25]	5	0.83	1.08	6.14	7.52	13.40	13.85	0	1.0	0	6.99	0.0	14.48
RPIVD [100]	200	1	1	5.46	117.63	85.0	989.0	0.0	1.0	0.0	118.08	0.0	1094.0
QFT [114]	0.8	0.0007	0.002	0.008	0.20	0.16	2.89	0.0	0.002	0.0	0.211	0.0	2.84
SimCLR+LSH [96]	0.04	0.001	0.009	0.035	0.11	0.16	0.47	0.0	0.005	0.0	0.098	0.0	0.55

terms of the evaluation of $ph_{similar}$ of images group (a&b) in Fig. 3.2. In addition, the perceptual hashing difference values for the altered images using groups (a&c) in Fig. 3.2 are represented too. The threshold T is directly brought from the original papers presenting those perceptual hash algorithms [25, 25, 78, 100, 114] except [96] that we generated. For a perceptual hash algorithm, if the perceptual hash value exceeds the threshold, an FNR is generated since the image is supposed to be authentic, but the algorithm detects otherwise. Whereas an FPR is generated when the image is supposed to be unauthentic, but the algorithm decides otherwise. Table 3.2 represents the outcomes of FNR and FPR for Facebook and Twitter platforms.

Looking at the average scores on Facebook evaluation at Table 3.1, it shows that all the results of tampered a&c are higher than similar a&b tests, which indicates that these algorithms recognize alteration. The selected T by the authors remains in between on all algorithms except Visual Model-Based (Vsul M-B) [25] that the average on a&b and a&c exceeded the T. Twitter, on the other hand, shows identical results on measuring (a&b) on all evaluations. These results show that the Twitter platform applies manipulation upon sharing in small factors value if the image meets the recommended dimension [68], which our selected dataset follows. For example, we upload an image to the Twitter platform with a size of 233KB and a dimension of 1280×720. After downloading the image back, we receive a new size, 230KB at the same dimension. Another trial was done that exceeded the recommended dimension on an image with a size 262KB and a dimension of 1850×1233. We received a new alteration with 82KB in size and 680×453 in

dimension. Hence, it is probable Twitter applies compression and re-scaling with different factors based on the image size and dimension.

Fig. 3.3 represents the percentage values of different image hashes a&c from Table 3.1 to similar image hashes a&b using Equation 3.3. If the gap percentage is high, it represents the algorithm's robustness in detection alteration. Otherwise, it is vulnerable for the algorithms to minimize the comparison score for similar tests and maximize the score for the altered tests. From the chart, the gap is suitable enough on the four algorithms: DCT , Wavelet , QFT, and SimCLR, regarding the Facebook metric with values of 7, 50, 25, and 3.14, respectively. The Visual Based algorithm kept the different gap on Facebook at the NT with a value of 6.23, which increases the model's sensitivity to distinguish between similarity and alteration. The other model, RPIVD went too low, even below the suggested T by the author. However, at the same time, NT shows a high gap, 25, and the highest accuracy amongst others. In contrast, on Twitter, as the similarity scores shows zeros at Table 3.1, the gap difference is significant where we solve the division by zero by adding a small fraction that a close to zero.

Finding optimal decision thresholds for social media platforms. The six perceptual hash algorithms under study were well designed by their authors with carefully-set decision thresholds in detecting image manipulation based on specific datasets [101, 119]. However, their authors determined the decision thresholds based on general modification/manipulation operation on images. In the specific social media platform scenario under study in this work, we want a perceptual hash algorithm to detect any deliberate image manipulation while at the same time not treating the image resizing/compression operation by the social media platform as image manipulation. Therefore, for each perceptual hash algorithm, there should exist a better decision threshold for the social media platform environment. In this section, we present the method for finding the optimal decision threshold and then verify that the performance will be better in comparison to those original decision thresholds as it appears at Table 3.3.

We recalculate the threshold for each algorithm based on the perceptual hashes outputs through the

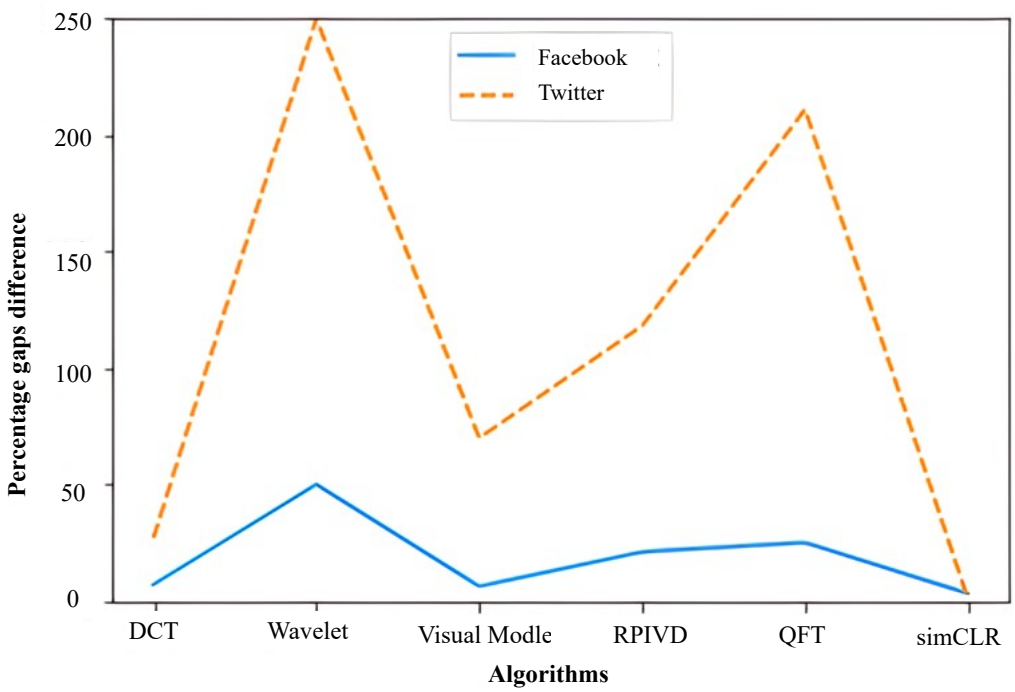


Figure 3.3: Perceptual hash gaps (*diff* defined in Equation 3.1) between similar images and tampered images.

conducted tests on Table 3.1. Based on [114], the optimal threshold can be determined using the probability of true authentication (PTA), which essentially calculates the probability distribution of $ph_{similar}$ and ph_{alter} results using TP and TN decisions of both social media platforms as it appears in Equation 3.5 and Equation 3.6. Based on the threshold value, which begins with zero, the probability of true authentication for similar images, i.e., $PTA_{similar}$, is represented by the blue line at Fig. 3.4 should be zero since all the images recognized as tampered; therefore, tampered images probability, i.e., $PTA_{tampered}$, shown by the orange dashed line should be one.

$$PTA_{similar} = \frac{\text{no. of TP results}}{\text{no. of similar image tests}} \tag{3.5}$$

$$PTA_{tampered} = \frac{\text{no. of TN results}}{\text{no. of tampered image tests}} \tag{3.6}$$

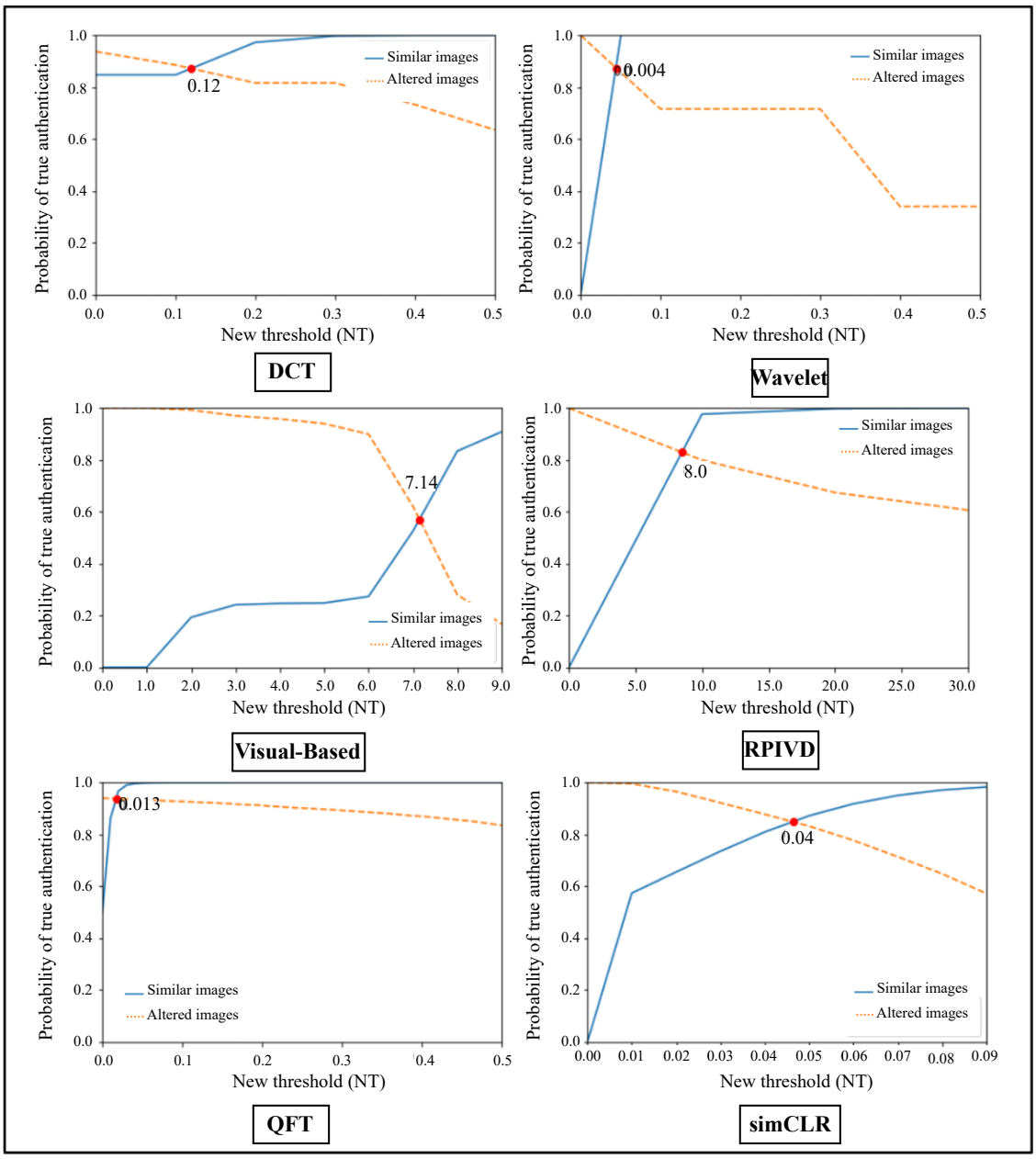


Figure 3.4: The new threshold (NT) calculation for each approach.

With each increase in the threshold value, the rate of true authentication for both states will change until they intersect at a certain point, as shown in Fig. 3.4, e.g., RPIVD with a red dot and value of 8. This point is selected to represent the new threshold (NT) where it balances the performance of FPR and FNR and gives the best accuracy. Table 3.2 shows the comparison of FNR and FPR rates for original thresholds (OT) and NT for all six algorithms. The outcomes of NT show

Table 3.2: FNR AND FPR comparison on scale of original thresholds (OT) and new thresholds (NT) for facebook and twitter.

Algorithm	Facebook						Twitter					
	Original Thresholds OT			New Thresholds NT			Original Thresholds OT			New Thresholds OT		
	OT	*FNR*	*FPR*	*NT*	*FNR*	*FPR*	*OT*	*FNR*	*FPR*	*NT*	*FNR*	*FPR*
DCT [25]	0.25	0.009	0.32	0.12	0.29	0.27	0.25	0.0	0.33	0.12	0.0	0.225
Wavelet [78]	0.004	0.013	0.63	0.004	0.013	0.63	0.004	0.0	0.64	0.004	0.0	0.64
Vsul M-B [25]	5	0.75	0.05	6.46	0.66	0.17	5	0.0	0.19	6.46	0.0	0.37
RPIVD [100]	200	0.0	0.79	8	0.099	0.098	200	0.0	0.79	8	0.0	0.15
QFT [114]	0.8	0.0	0.92	0.013	0.16	0.14	0.8	0.0	0.9	0.013	0.0	0.19
SimCLR+LSH [96]	NA	NA	NA	0.04	0.16	0.17	NA	NA	NA	0.04	0.0	0.26

Table 3.3: Facebook and Twitter accuracy comparison between original thresholds (OT) and new thresholds (NT).

Algorithm	Facebook				Twitter			
	OT	*Accuracy*	*NT*	*Accuracy*	*OT*	*Accuracy*	*NT*	*Accuracy*
DCT [25]	0.25	75.33%	0.12	71.34	0.25	75.24%	0.12	85.47%
Wavelet [78]	0.004	67.50%	0.004	67.50	0.004	67.65%	0.004	67.65%
Vsul M-B [25]	5	59.42%	6.46	58.42	5	87.71%	6.46	81.21%
RPIVD [100]	200	60.33%	8	90.12	200	60.40%	8	92.06%
QFT [114]	0.8	53.50%	0.013	84.21	0.8	53.67%	0.013	90.48%
SimCLR+LSH [96]	NA	NA	0.04	82.87	NA	NA	0.04	86.70%

significant enhancements on Facebook and Twitter platforms with minor changes at Wavelet and Vsul M-B algorithms. Accuracy comparisons at Table 3.3 represents remarkable changes too.

We conducted an experiment on SimCLR for Facebook images that shows a promising approach. We used a Pretrained SimCLR as feature extraction. Then, We replaced both LSH and hamming distance with a Normalized Euclidean Distance as shown in Equation 3.7. Also, we fed the output of the learned representation, which is the output of the final convolution layer to the Normalized Euclidean Distance instead of the output of the fully connected layer as suggested in [96]. The accuracy of the model increased to 90%. However, the output of SimCLR representation is a 4069 length vector where each element of the vector is represented by 32-Floating Point bits. This experiment suggests that SimCLR results have the potential to be improved.

$$d(u,v) = \frac{1}{2}\frac{\sigma^2(u-v)}{\sigma^2(u)+\sigma^2(v)+\epsilon} \tag{3.7}$$

The distance denoted by *d* measures the normalized euclidean distance between two images vectors *u* and *v*. The first term in the numerator calculates the difference between the two image vectors and their variance afterward. The denominator term refers to the summation between the two vectors' variance and ϵ, which is a small number added for numerical stability. The result of the quotient, then, is multiplied by 1/2.

The average performance analyses are shown at Table 3.4. It is clear that increasing operation for feature extraction also increases overhead. [25] and [78] were the lightest overall on feature processing whereas [96] time consumption is the highest by a significant time. Therefore, the algorithms of images systems should consider different types of environments instead of using highly advanced resources in the developments pipeline.

Table 3.4: Average performance analyses.

Algorithm	***Processing Time (s)***
DCT [25]	0.023
Wavelet [78]	0.065
Vsul M-B [25]	1.76
RPIVD [100]	0.18
QFT [114]	0.087
SimCLR-LSH [96]	11.76

3.3 Discussion

Image feature preservation plays the primary role in designing a robust authentication system. The six approaches follow different ways of vector preservation and apply one or more operations for that objective, such as RPIVD applying four statistical measures (mean, variance, skewness, and kurtosis). These operations could generate a unique perceptual hash. However, it could increase performance overhead.

Selecting the optimal threshold for making a decision is challenging. The authors of those six evaluated algorithms determine the threshold based on their targeting image alteration methods and the dataset used in training or evaluation. For instance, most of these algorithms target many types of image attacks. However, at the same time, they focus on one alteration, such as [78] for image compression and [100] for image rotation. A study of image authentication with minor alteration could be an area of interest with high security, considering lowering the collision rate and minimizing the threshold T. A theory of cryptographic hashing in uniqueness and identical could be targeted in perceptual hashing developments.

The hash value size is a significant parameter in finding a solid and efficient perceptual hash algorithm. Increasing the hash size can improve the accuracy of the decision, but at the same time, the computational cost of the hash and the storage requirement will also increase. The future work on this aspect is to find the suitable trade-off between accuracy and efficiency in image authentication for different applications.

We also intend to investigate more about the machine learning approach. The initial results show that SimCLR with LSH is a promising approach to image authentication. However, there is a place for improvement. We can design a custom contrastive learning that serves the image authentication task to enhance the results. Further, we can conduct more tests on SimCLR model with Normalized Euclidean Distance to validate and improve the accuracy and model performance. Also, we can look into other neural network architectures that may work in the image authentication domain.

CHAPTER 4: IMAGE AUTHENTICATION USING SELF-SUPERVISED LEARNING TO DETECT MANIPULATION OVER SOCIAL NETWORK PLATFORMS

Building on the previous chapter, our goal in this chapter is to focus on improving the authentication process. We start by explaining our approach and how it addresses the potential attacks. Next, we demonstrate our findings based on the data collected in our previous work, SMPI. Finally, a discussion statements is introduced.[1]

4.1 Methodology

In this section, we first discuss the assumptions to authenticate images over the social media platforms used in our research. Then, we explain the details of our proposed new system.

Application Scenario and Assumptions. To ensure security, the following assumptions are proposed. It is assumed that users must create accounts using their valid information to use our system. Naturally, Twitter accounts holders are verified by Twitter, Inc. The generated pHash of an image from our system can be added to the image through the description feature on Twitter's platform in hexadecimal representation (e.g., see Fig. 4.1). End-users can download said image, copy its pHash from the description, and use our system for authentication. Moreover, users can re-publish the image with its pHash on their account. Adversaries here are forced to provide their credential information to our system in order to generate a new pHash. This assumption applies on other social media platforms, e.g., Facebook.

Proposed System. As shown in 4.2 and Algorithm 1, the system design consists of four stages: pre-processing, feature extractor, contrastive loss, and pHash generator, respectively. We describe each stage below:

[1]The contents of this Chapter are based on our publication to IEEE MILCOM 2022 [2]

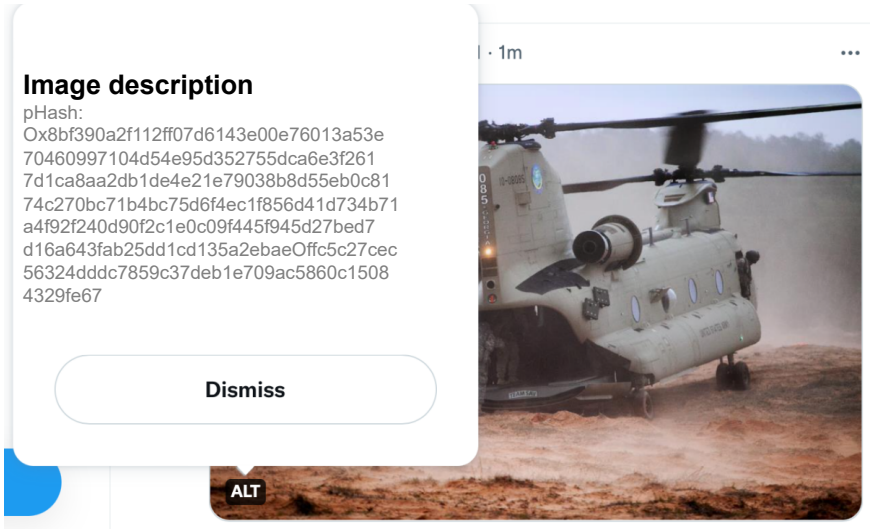


Figure 4.1: Sample image posted on Twitter platform that has pHash on the Image description feature.

Source: CASIA dataset.

1) pre-processing: Before each image passes to the training phase, it goes through data augmentation, resizing, applying random color jitters, and random compression. Data augmentation improves performance when applied to deep learning models, as is shown in SimCLR [17]. SimCLR is a self-supervised learning model that uses data augmentation to generate two augmented images of each image in the batch and minimize the difference during the training task. Our model uses SimCLR architecture with crucial modification. We introduce a new step to the augmentation process to suit our target task. Instead of creating only two augmented images as it has been done in the original SimCLR, we, also, create a content-changing sample from the original image x called an *altered* image $\tilde{x}_{alt}$. The alteration is added to the image randomly selected from one of three image modification techniques *copy-move*, *splicing*, and *removal*. The copy-move ($cp - mv$) alteration is an operation of randomly copying a spot of an image with size of m×m×3, where m $\in \{16{:}208\}$ and randomly pasting it on a different location of the same image. The splicing (sp) is the same process as copy-move in randomization, but the spot is pasted on a different image. Finally, the removal (rm) alteration is where we follow the same technique of selecting a spot with a random size and a random location, but we apply kernel simple blurring filter 50 times on the

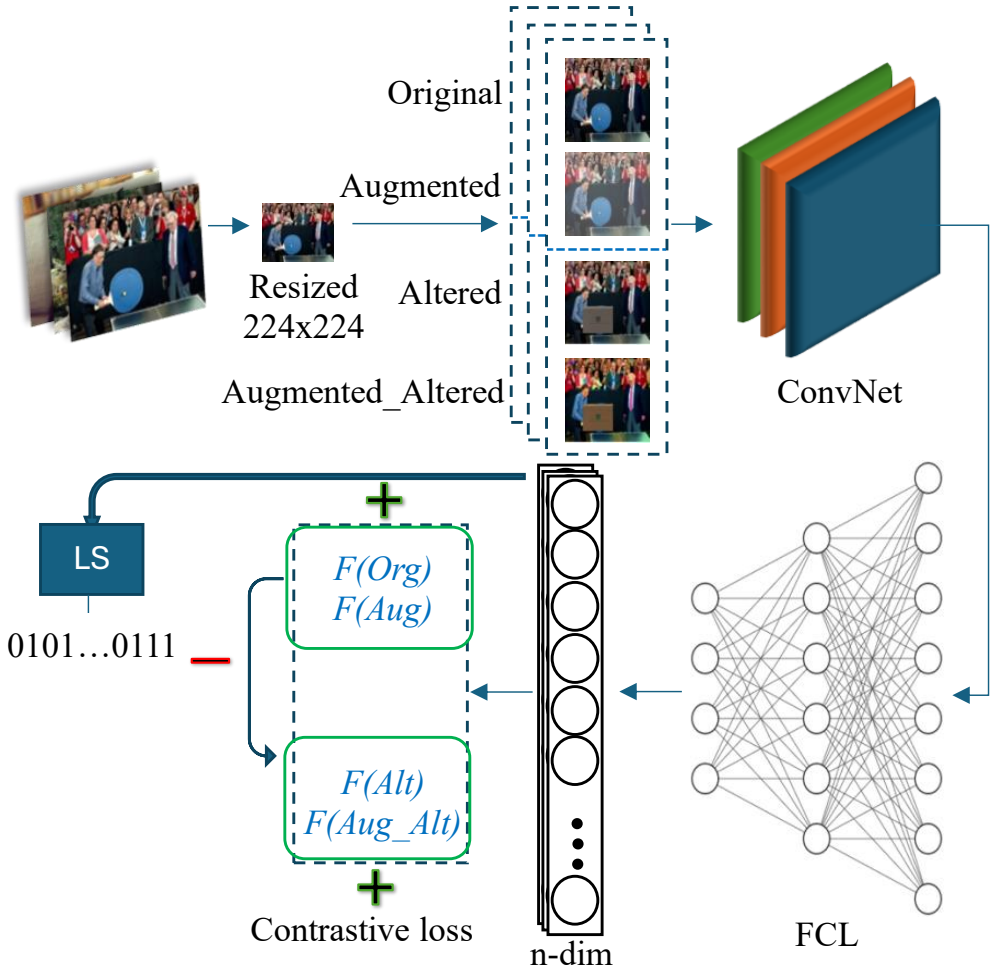


Figure 4.2: Proposed approach for image authentication.

same spot without moving it to a different location.

As in Fig. 4.2, the augmentation process in this approach applied to original and altered images to cover multiple distorted versions. From these augmentations and alterations, each image is converted into two pairs after fixed resizing to 224×224×3: original x with random *augmented* $\tilde{x}$ and *altered* x_{alt} with random *augmented_altered* $\tilde{x}_{alt}$. The augmented refers to the version of the original image with content-preserving manipulation. On the other hand, altered represents the content-changing manipulation. Finally, augmented_altered is the copy of the altered image with content-preserving manipulation. The next stage in the training knows that each pair is authentic

on its own and unauthentic in comparison to the other pair.

2) feature extractor: Each image is passed to a convolution neural network (ConvNet) as shown in Fig. 4.2. ConvNet produces feature maps that capture image features. Next, these feature maps are flattened and mapped to a n-dimensional (n-dim) feature vector through a fully connected layer (FCL), (z, $\tilde{z}$, z_{alt}, and $\tilde{z}_{alt}$), where n-dim is a hyperparameter representing the number of nodes in the last layer chosen ahead of the training. The ConvNet used is ResNet-18 which consists of convolution layers, poling, ReLu, and skip connection.

3) Contrastive loss: At the training stage, the n-dim vector is passed to the loss. We use contrastive loss as used in [17] to maximize the agreement between the positive samples and minimize the agreement between the negative samples by minimizing the normalized temperature-scaled cross entropy loss (NT-Xent). We assign the temperature τ to 0.1 as suggested by [17]. SimCLR uses the augmented pair as positive samples. Negative samples are collected by pairing an image with another in the same batch that is not its augmented twin. However, our proposed approach takes the original image and its augmentation as a positive example, as well as the altered version with its augmentation. The negative sampling is the same as the original approach in SimCLR. Consequently, the original image will be paired with its altered version as a negative example to force the model to distinguish between images that shares high level features with content modification.

4) pHash generator: At the evaluation step, LSH is used in the hash generation to convert the long length of extracted features into small binary bits representations by mapping close feature vectors to buckets with similar hash values [32, 94]. Random projection is one type of LSH we used because it provides an independent secret key during random hyperplane generation that can be adaptive for security purposes. In practice, the feature vectors from FCL are flouting points with a length of 512 multiplied by a random hyperplane matrix of the size of 1024x512. This matrix multiplication is finally converted to a bit vector by applying a Heaviside step function to each element. The final generated hash length is 1024 bits.

Algorithm 1 Training Stage for the Proposed Model.

Training Stage:

input: batch size $N \times 2$, constant τ

network: ResNet-18 (f) + FCL (g)

for `randomly sample` $x \in \{X\}$ **do**

draw one attack $a \in \{cp - mv, spl, rm\}$

$x_{alt} = a(x)$

draw two augmentation functions $d \sim aug$, $\tilde{d} \sim aug$

$\tilde{x} = d(x)$

$\tilde{x}_{alt} = \tilde{d}(x_{alt})$

Forward Pass:

$z = g(f(x))$

$\tilde{z} = g(f(\tilde{x}))$

$z_{alt} = g(f(x_{alt}))$

$\tilde{z}_{alt} = g(f(\tilde{x}_{alt}))$

end for

Calculate contrastive Loss L

update the network parameters to minimize L

return network $f(.)$ and $g(.)$

Evalution Stage:

input: x, $\tilde{x}$, constant random matrix r (1024×512), and Threshold θ

output: 'Similar' if d $\leq \theta$ else 'Altered'

$e = g(f(x))$ #feature vectors for img1

$\tilde{e} = g(f(\tilde{x}))$ #feature vectors for img2

$p1 = (e \cdot r^T \geq 0)$ #binary hash for img1

$p2 = (\tilde{e} \cdot r^T \geq 0)$ # binary hash for img2

$d = HammingDistance(p1, p2)$

4.2 Experimental Results

This section explores the experiment setup, the training configuration, and the main results based on F1-score metric.

Experimental Setup. To evaluate the robustness and effectiveness of the proposed scheme, we run a large number of experiments. Our implementation and training were done using NVIDIA

Table 4.1: The structure of training and validation sets.

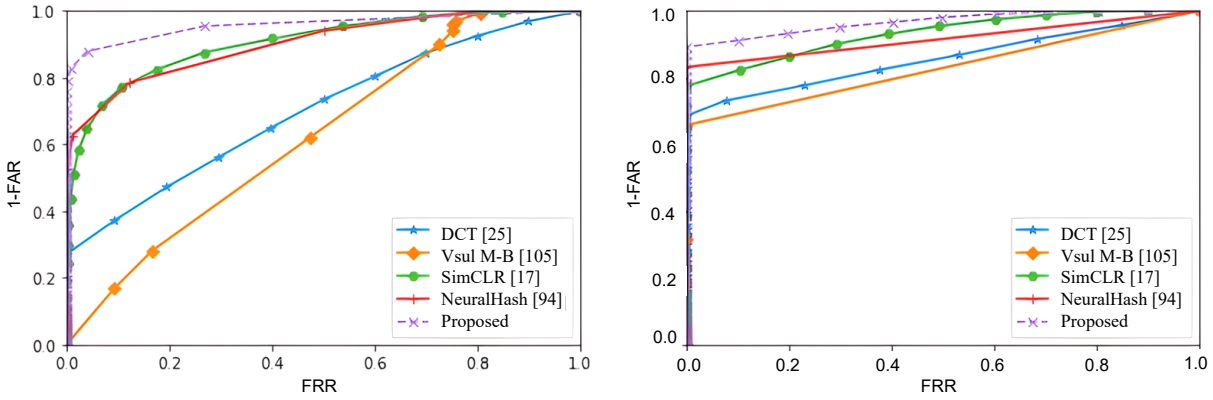
Stage	Dataset	no.
Training set	Flickr [41]	8,000
	Holopix50k [43]	41,000
	Tiny ImageNet [52]	100,000
	PS-Battles [39]	10,000
	ImageCLEF [65]	21,000
Validation set	SMPI [5]	19,458
	IMD2020 [67]	200
	COVERAGE [110]	200

GeForce RTX 3090 GPU. All other prior models were re-implemented and tested using the Colab platform based on best effort resources. We use PyTorch-lightning[2] open-source python library for our proposed system. Moreover, we re-implemented or used provided sources of other schemes and integrated a final hash generation using LSH for deep learning approaches.

Training Configuration. The configuration of the training goes through multiple processes. First, we collected 180,000 images from different resources, as shown in Table 4.1. This diversity prevents bias to any image classes. Next, four modified samples were derived from each original image sample after resizing into $224\times224\times3$ and paired into two groups. The first pair contains the original image and its augmented copy, and the second contains the altered and its augmented altered sample. Thus, the total number of training examples was increased to reach 720,000 images. We trained our model with contrastive loss to increase the agreement of similar images and decrease the agreement of dissimilar images. We used SMPI [5] as a benchmark for evaluating images that were collected from Facebook and Twitter platforms. In addition, we tested our model on the datasets IMD2020 [67] and COVERAGE [110] to compare it with other state-of-the-art models.

Main Results. The similarity metric we used for our approach is Hamming Distance measurement, as used by [17, 25, 94]. The algorithm presented in [105] is the only one that used Euclidean

[2]https://www.pytorchlightning.ai/



(a) Facebook.

(b) Twitter.

Figure 4.3: Comparison of ROC curves for each authentication model using SMPI dataset.

distance. The pHash distance d between two images draws the line of the threshold θ that will be the indicator in our image authentication system. For similar images, the distance d should satisfy $0 \leq d \leq \theta$. For altered or dissimilar images d should be above the threshold θ, , i.e., $\theta < d$. Table 4.2 shows the best selected θ based on the best F1-score assessments.

Table 4.2 compares five schemes based on the SMPI dataset. The bold F1-scores in each column are the best-reported score, which shows the significant improvement using our proposed approach on both social media platforms. Overall, four models [17,25,94,105] have close F1-score at Twitter with 0.87, 0.84, 0.88, and 0.88 respectively. Our proposed scheme reached the highest score by 0.99. In contrast, [17, 25, 94, 105] under-perform with Facebook with 0.70, 0.44, 0.82, and 0.82 respectively and a new high record achievement with the proposed model by 0.92.

We evaluated the performance of the models using the Receiver Operating Characteristic (ROC) curve as illustrates at Fig. 4.3. The X-axis is the probability of False-Reject Rate (FRR), the ordinate is the probability of False-Accept Rate subtracted from one $(1 - FAR)$. An ROC curve that is closer to the top left corner means a better performance of content authentication. From ROC curves of the five schemes in Fig. 4.3(a), we can observe that our scheme achieves the best ROC curve on the Facebook platform compared with the others, whereas model in [105] is the

Table 4.2: F-score results of each model using SMPI dataset and their threshold NT.

Model	NT	**Facebook**	**Twitter**
DCT [25]	0.12	0.70	0.87
VisualModelBased [105]	6.46	0.44	0.84
SimCLR [17]	0.04	0.82	0.88
NuralHash [94]	0.02	0.82	0.88
Proposed	0.02	**0.92**	**0.99**

Table 4.3: Area under curve (AUC) performance comparison.

Model	**IMD2020 [67]**	**COVERAGE [110]**
CFA1 [30]	0.586	0.485
J-LSTM [11]	0.487	0.614
ManTra-Net [111]	0.748	0.819
TraFor-Self [36]	0.848	0.884
Proposed	**0.98**	**0.99**

worst. Fig. 4.3(b) represents the ROC curve on the Twitter scale shows small gaps in most models, where our proposed technique overpasses the others.

Moreover, we evaluated the proposed algorithm using other datasets IMD2020 [67] and COVERAGE [110] that are mainly founded for image forgery assessing on the scale of copy-move, splicing, and removal. We picked real-life manipulated images part from [67] that are collected from the Internet. Based on the same θ that we picked previously for our model, the Area Under Curve (AUC) performance comparisons is provided at Table 4.3 with other models [11,30,36,111]. Our model's results, which are in bold, are ahead of others by a high percentage.

4.3 Discussion

Many magnificent works used deep learning for image classifications and were based on a large-scale dataset such as [24]. On the other hand, image authentication received less attraction due to multiple reasons. First, most image datasets are generated with big alterations; therefore, many developed systems accomplished high accuracy on those datasets. second, small alteration to the

image is hard for the systems to detect, and the concept of pHash authentic distorted copies of the original image. Third, distorted copies of the original image have unlimited and unknown factors. For instance, the compression quality factor during the model design is fixed, i.e., quality factor $\in \{1, 5, 10, 30, 50, 70, 90, 95\}$ and the tested images are compressed with random and unknown values 1:100. This example is one case where nine other image operations can be implemented on an image and considered as an authentic copy, not counting that one image might receive multiple operations.

For instance, Facebook deals with each user differently during exploring the platform because Facebook needs to allow their users with weak network coverage to use their platform by applying different compression quality factors based on the network status. For example, we examined downloading a shared image from the same post by two PCs, we received different sizes of the same image. This alteration itself makes the task of authentication complex.

In addition, the length of pHash plays a significant part in image authentication. The larger is better to make the extracted feature vectors more sensitive. In contrast, the larger length would cause more overhead on the payload of the image in practice. Therefore, we remain our evaluation on 1024-bits to make it more applicable to various applications.

Finally. we looked into another direction to enhance our proposed model. We increased the number of altered and augmented version of each image to cover more samples in the batch. Therefore, the batch consists mostly of one image and its n-altered and n-augmented versions. The results showed degrading in the performance of the model. We concluded that increasing the samples of the original image in the same batch harm the model efficiency. Therefore, we limited the model to have only four samples from the original image. These samples are original, augmented version, altered and augmented of the altered version.

CHAPTER 5: ADVERSARIAL-AWARE DEEP LEARNING SYSTEM BASED ON A SECONDARY CLASSICAL MACHINE LEARNING VERIFICATION APPROACH

This chapter explores the most common adversarial attacks that target deep learning models, particularly in the field of computer vision. We deliver our motivation, analyze the adversarial attacks, define their common ground, and present a new defense technique based on a secondary verification approach that is not affected by these attacks. Our defense solution is inspired by the beginning of satellite communication. We also provide suggestions to increase security strength and give application developers the freedom to choose their own security options. Furthermore, we compare our approach with others and present the results, and finally introduce our discussion.[1]

5.1 Motivation and Threat Model

Motivation. After multiple assessments of the different adversarial attacks on different DNN models, we notice that once the attack succeeds on one deep learning model, it succeeds on other models as well, as shown in Table 5.1, which was obtained by running multiple adversarial attacks (FGSM, Deepfool, CW, and PGD) on ResNet-34 [38] as a target model using the CIFAR-100 dataset. The generated adversarial samples are then tested on VGG16 [90] and DenseNet [44] DNN model classifications. We find that the accuracy of the targeted model is similar to that of untargeted DNN models. Researchers in [108] addressed the same issue, naming it "transferability" of adversarial examples, meaning that the generated samples from adversarial attacks on one targeted DNN model may work on different untargeted DNN models. Therefore, a model obtained by a different approach is interesting to study, and we selected a random forest (RF) [56] decision-tree-based classifier model for our study, considering all the challenges of using this limited model for image classification.

[1]The contents of this Chapter are based on our publication to Sensors 2023 [7]

Table 5.1: Accuracy comparison of different DNN models before and after adversarial attacks on the CIFAR-100 dataset.

Attacks	**Targeted model**	**Un-targeted models**		
	ResNet-34	VGG16	DenseNet	
w/o attack	77.47	72.25	78.69	**Accuracy(%)**
FGSM [33]	34.25	35.09	36.19	
Deepfool [64]	25.78	24.79	24.84	
CW [15]	25.77	24.49	25.0	
PGD [60]	22.58	22.87	22.7	

Threat model. Our threat model assumes that the attacker knows there is a detection method employed but does not know what it is. In this setting, only the DNN model and its parameters are known to the adversary.

5.2 Methodologies

We introduce our proposed adversarial attack detection method in this section. Our primary image classification system, shown in Figure 5.1, is based on the DNN approach, and we choose ResNet with 34 layers here for our investigation. The primary model could be any other DNN model that uses backpropagation because adversarial attacks exploit backpropagation to optimize the perturbations introduced to the input data on DNN models. The input is an image that could be a real image with no alteration or an adversarial generated sample from one of four attacks: FGSM, Deepfool, CW, or PGD. Our output of ResNet-34 is the highest probability index that indicates the class the image belongs to, which is referred to as Top_1 classification.

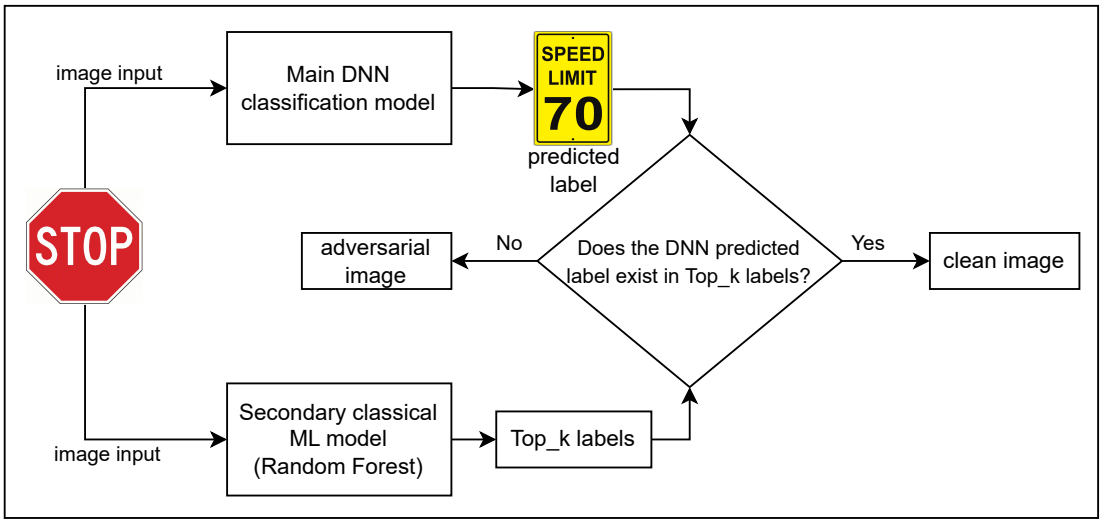


Figure 5.1: Proposed adversarial detection system design, which is composed of a primary DNN classification decision model and a secondary classical ML model for adversarial attack detection and verification.

Unlike the primary approach, we use the classical ML model, the random forest (RF) model, as a secondary model for adversarial attack detection. The model can be vulnerable to adversarial attacks, as seen in [10], with the aim of deceiving the intrusion detection system. Nonetheless, we opt to use this as a secondary model because it employs a different method and does not rely on the gradient technique utilized in computer vision adversarial attacks. Therefore, the perturbation added to the images does not impact the model's image features or the model classification performance.

The RF model is a decision tree module based used in regression and multiclassification problems [9, 22, 74, 91, 93]. It is an extension of the bagging method, as it utilizes both bagging and random feature selection to create an uncorrelated forest of decision trees. It also reduces overfitting and increases the diversity of the trees in the forest. The randomness in selecting the features for each tree determines and eliminates the inserted perturbations information on the adversarial samples, as illustrated in Figure 5.2, where the accuracies of RF model before and after different attacks are almost identical. In the same figure, the k-NN model is demonstrated as a classical ML

model that is not affected by the added perturbations as well.

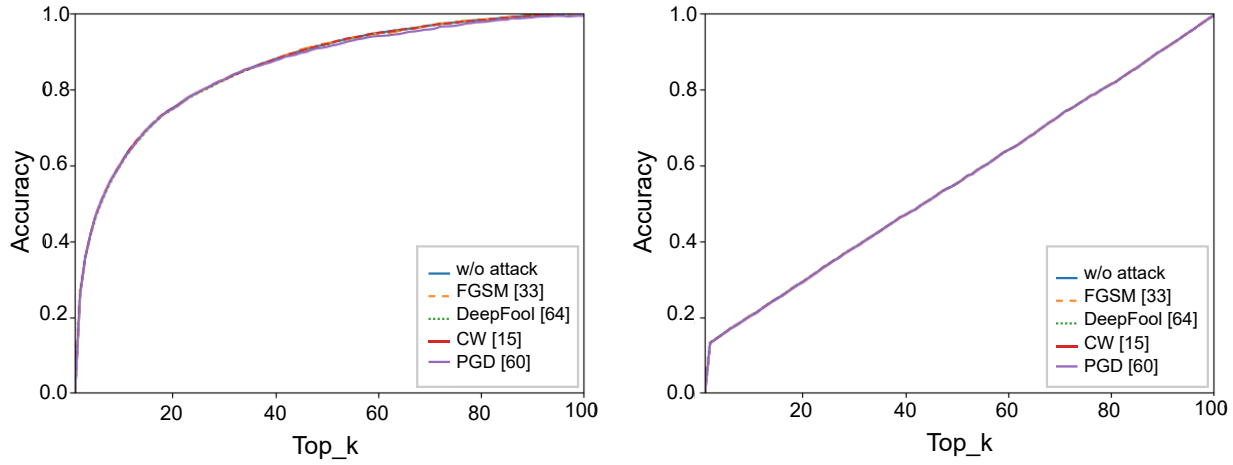


Figure 5.2: Classification accuracy over Top_k before and after different adversarial attacks using the CIFAR-100 dataset by two classical ML models: (**a**) random forest model and (**b**) the k-NN model. The accuracies under different adversarial attacks are almost identical; thus, those resulting curves override each other and make a single purple-colored curve.

Our outputs of RF are the top k indices (Top_k) of the predicted class probabilities for the inputs. We selected Top_k and relied on it for our study to match the accuracy of the RF model with the selected DNN model on the CIFAR-100 dataset, which has 100 classes. Top_1 in the RF represents the worst accuracy, as illustrated in Figure 5.2, whereas Top_100 represents 100 percent accuracy because its decision is always correct, where the decision group includes all possible classes. When the k parameter in Top_k equals 22, the accuracy reaches around 77 percent, the same percentage as the primary DNN method prediction accuracy in the top_1 classification. Moreover, by adjusting the value of k in the Top_k classification, our methodology provides more control to its users and more choices to select optimal security versus classification accuracy based on the AI application design, as described in depth in Section 5.2.

Category of Image Dataset. Under adversarial machine learning (AML), we run each adversarial attack individually on the DNN model, ResNet-34, using the test set in the CIFAR-100 dataset, which contains 10K images. The attack success ratio of each of the adversarial attacks varies, as illustrated in Table 5.1. During the categorization process, as represented in Algorithm 2, each

image (x) is first checked by the DNN model for the correct label. The mispredicted result from DNN(x) adds x directly to the SETmis set. In contrast, the correct prediction of x passes to the AML(x) algorithm for a trial (e.g., FGSM), and the successfully applied attack output is added to the SETadv set. The unsuccessful attack moves x to the SETcrc set. In summary, we categorize the outputs into three sets as follows:

- SETcrc: The set of images that the DNN can correctly identify;
- SETmis: The set of images that the DNN misidentifies (misclassification);
- SETadv: The set of images produced by AML that can successfully and deliberately make the DNN misidentify as another object the attacker wants.

The percentage of misclassified images (SETmis) is maintained at 22.54 across various attacks. However, the percentages for the other categories vary depending on the strength of each attack and its parameters. Generally, the adversarial generated samples (SETadv) or the attack success ratio receive the highest percentage among other sets in all four adversarial attacks.

Detection Algorithm. The adversarial image detection model, denoted as $Adv - aware(x)$, is addressed in Algorithm 3. We first pass a test image (x) to the primary DNN model, which is DNN(x) with Top_1, and to the secondary model, which is the RF model with Top_k donated, as denoted by RF(x, k). Then, we have two outputs: a single-class prediction from the primary DNN model (y) and k class predictions from the secondary model (Top_k) as a list of k classes. We check whether y predicted classes exist in the Top_k prediction list. If y exists in the Top_k, then it returns a boolean "false" value for forged status with the DNN(x) label (y). Otherwise, it returns "true" without a label or none, which detects a possible adversarial sample.

For instance, as shown in Figure 5.1, we use a *STOP* road sign as an input sample to our model. It passes to the primary model and the secondary model concurrently. In an adversarial attack scenario where the *STOP* sign image is a manipulated image, the predicted class from the primary model is *SPEED LIMIT 70*, whereas the second model provides a Top_3 list of predictions,

Algorithm 2 Categorize Image Dataset.

[SETcrc, SETmis, SETadv] = Category (Image dataset, DNN classification results)

Input: $\{x,\ right\,label\} \in$ CIFAR-100$(test\ set), DNN\ model\ DNN(x),\ adv_attack\ AML(x)$

Output: The three categories of image dataset according to DNN model classification and AML results.

Initialize SETcrc, SETmis, SETadv to be all empty

for image $x \in$ CIFAR-100 **do**

  if $DNN(x)$ is mispredict **then**

    SETmis $\leftarrow$ x

  else

    if $DNN(x)$ is correct and AML(x) fail **then**

      SETcrc $\leftarrow$ x

    else

      SETadv $\leftarrow$ x

    end if

  end if

end for

return [SETcrc, SETmis, SETadv]

for example, *STOP*, *Roundabout*, and *No entry*. Our model detects the input image as a forged "true", since the predicted class from the primary model does not exist in the list of the secondary model. In the case of clean detection, the predictions have to be found in both model predictions. During our evaluation, we excluded misclassification samples in this section, which are tackled in Section 5.2.

Defense System Adaptive Design. This section discusses a new technique for selecting the best value of k in the Top_k used in the secondary model based on the underlying application-specific requirements in terms of accuracy and security. Some applications require zero tolerance for attack success. On the other hand, a low success ratio of adversarial attacks in some other applications does not cause severe damage. Moreover, including the misclassification samples in this adaptive design improves the overall detection accuracy of adversarial attacks. The details are explained in the following subsections.

Algorithm 3 Adversarial-Aware Deep Learning System.

[forged, label] = Adv-aware (x)

Input: image x.

Output: Whether the image is forged by adversarial attack or a clean image; classification label if x is a clean image.

$y \leftarrow$ DNN(x) # DNN model classification label for the image x

Top_k $\leftarrow$ RF(x, k) # The top k group of labels generated by the RF classification model

if $y \in Top\_k$ **then**

    forged = false; label = y

else

    forged = true; label = None

end if

return [forged, label]

Outputs of Our Proposed Adversarial-Aware Image Recognition System. Our image recognition system has two possible outputs: (1) the image under inspection is authentic, and its *identified* label is provided, or (2) the image under inspection is forged by AML and tagged as *forged*. Therefore, given that there are three possible sets of images in terms of DNN identification (introduced in Section 5.2), here are the six possible decision scenarios for our proposed system:

- Decision A (Dec_a): An image in SETcrc that is authentic and correctly identified;
- Decision B (Dec_b): An image in SETmis that is correctly identified as forged;
- Decision C (Dec_c): An image in SETadv that is correctly identified as forged;
- Decision D (Dec_d): An image in SETcrc that is misidentified as forged;
- Decision E (Dec_e): An image in SETmis that is misidentified as authentic and misclassified;
- Decision F (Dec_f): An image in SETadv that is misidentified as authentic.

From a user's perspective, Dec_a, Dec_b, and Dec_c are all 'good' and rightful decisions, whereas Dec_d, Dec_e, and Dec_f are wrongful decisions that could cause a negative impact/cost to the user.

Adjustable Parameter in Our Proposed System. In our proposed adversarial-aware image recognition system, one critical parameter that can be adjusted/controlled by the end user is the value of k in the Top_k classification by the secondary model. It can be used to make a delicate tradeoff between increasing the defense accuracy of adversarial attack images and increasing the correct recognition of normal images. The secondary verification of the ML module determines if an image under inspection belongs to one of the Top_k classes among all possible classification classes. Its classification setting (Top_k) can be, for example, Top_1, Top_10, Top_20, etc. When k increases, the classification decision by the DNN module has a higher probability of being included in the Top_k classes of the secondary verification system, which increases the possibility of good Dec_a and the possibility of bad decisions (Dec_e and Dec_f) as well.

In this work, we present a solution to the above dilemma by translating and quantifying the problem into the optimization of a carefully defined objective cost function. We explain it in detail below.

Using Objective Cost Function to Achieve Optimal Defense. Generally speaking, in most computer vision applications, a successful AML attack causes much more damage to the user than a misclassified event. In most cases, misclassifying an object/content in an image leads to a clearly identifiable wrongful conclusion, such that the user can easily know that it is a wrong identification, for example, misidentifying a road STOP sign as a red balloon in autonomous vehicle driving indicates that this is wrong image identification. However, a successful AML attack could make the user misidentify the STOP sign as a SPEED LIMIT sign, which could result in a serious car accident.

For this reason, when we decide how to adjust detection and defense settings for our proposed system, we should not use the classification accuracy, AUC score, or attack success rate directly as the metric. Instead, we define an overall cost objective function, that is, the weighted summation of all image classification results, to find the optimal defense parameters that minimize this objective function.

For the six decision outputs of our proposed system (Dec_a to Dec_f), each decision for one image

has its own cost (due to misidentification) or gain (due to correct identification), which can be treated as a positive or a negative cost. Let us define C_a, C_b, and C_c as the gains for each of those three good decisions (Dec_a, Dec_b, and Dec_c) and C_d, C_e, and C_f as the cost values for each of those three wrongful decisions (Dec_d, Dec_e, and Dec_f).

The objective cost function ($Objf(k)$) for choosing the optimal defense parameter (Top_k) in the secondary RF classification module is illustrated in Algorithm 4 and shown in Equation (5.1). We find the optimal value of k by selecting the minimum output (min_k) from the equation when changing k from 1 to 100. Parameters N_a to N_f refer to the number of times when decisions Dec_a to Dec_f happen, respectively.

$$Objf(k) = min_k(C_d \cdot N_d + C_e \cdot N_e + C_f \cdot N_f - C_a \cdot N_a - C_b \cdot N_b - C_c \cdot N_c) \tag{5.1}$$

To calculate N_a, N_b, ..., and N_f, a loop is conducted over the entire test set of the CIFAR-100 dataset. In Algorithm 4, each image (x) from the dataset is previously divided into three sets by Algorithm 2 (SETcrc, SETmis, and SETadv). Each if statement checks whether x image belongs to one of the sets and whether the outcomes of each model prediction (DNN and RF) are matched. For example, suppose x is a human object and DNN identifies it correctly, and the prediction also exists in the Top_3 RF outcomes. In that case, the decision state is set to Dec_a and N_a counter increments by one.

This optimization is conducted after the training stage, when we know the ground truth of all images, as shown in Section 5.3, and can calculate the values of N_a to N_f for each Top_k parameter for all test images. Since the number of possible values of k is limited (in our model, it has 100 possible values ranging from 1 to 100), there is no technical challenge in solving the optimization problem.

Examples of Adjusting Weights on different applications. In this section, we use several application scenarios to show why they need different cost weights in our adaptive design and the above

Algorithm 4 Adaptive Design Algorithm.

[k] = Adaptive(DNN classification results, RF classification results)
Input: CIFAR-100(test set), DNN, RF
Output: optimal parameter k for the secondary RF model
for $k \in \{1, 100\}$ **do**
  Set all the counters $N_a, N_b, \ldots, N_f$ to 0
  for image $x \in$ CIFAR-100 **do**
    if $x \in SETcrc$ & $DNN(x) \in RF(x, k)$ **then**
      N_a++
    else
      N_d++
    end if
    if $x \in SETmis$ & $DNN(x) \notin RF(x, k)$ **then**
      N_b++
    else
      N_e++
    end if
    if $x \in SETadv$ & $DNN(x) \notin RF(x, k)$ **then**
      N_c++
    else
      N_f++
    end if
  end for
  Objective function $f(k) = (C_d \cdot N_d + C_e \cdot N_e + C_f \cdot N_f - C_a \cdot N_a - C_b \cdot N_b - C_c \cdot N_c)$
end for
Among all $f(k), k \in \{1, 100\}$ find the minimum $f(k^*)$
Return the optimal index k^*

optimization Equation (5.1). In different image classification applications, users can define the concrete values for the other cost factors according to their expert opinion and application scenarios. Four applications are introduced in the following, and the next section presents the outcomes of this adaptive method.

- **Self driving**: We can define C_a = 0.3, C_b = 0.1, and C_c = 0.5. The value of C_c is higher than C_a because in self driving, it is more important for us to detect an adversarial attack than to correctly identify a normal roadside sign image. Similarly, we can define C_d = 0.1, C_e = 0.3, and C_f = 0.8. We define C_f as having a significantly higher value than others because Dec_f means autonomous driving is compromised under a deliberate adversarial

attack. For example, we could treat a STOP sign image as a right-turn-only sign, which could result in serious accident consequences. The value of C_e is higher than C_b in detecting misclassified images by the model due to the risk value we assume.

- **Healthcare**: Although deep-learning-based healthcare systems could achieve high accuracy in disease diagnosis, few such systems have been deployed in highly automated disease screening settings due to a lack of trust. Therefore, the human-based double-check process is usually used, and hence, the deep learning healthcare system can be tolerated in the security. Example values of the weights are C_a = 0.7, C_b = 0.4, C_c = 0.1, C_d = 0.4, C_e = 0.1, and C_f = 0.3. C_a is the highest cost weight because the physician will most likely discover failure in other decisions during manual double checking.

- **Face recognition in checking work attendance**: Misrecognition or adversarial impact is low because the potential of utilizing these challenges by the employees is rare. Therefore, we can obtain higher positive gain values with C_a = 0.7, C_b = 0.4, and C_c = 0.2. In contrast, we can value the negative decisions as C_d = 0.4, C_e = 0.2, and C_f = 0.2.

- **Detecting improper digital content**: Mispredicting nudity images to protect children is another example where the costs of an AML attack are medium—not as risky as in autonomous driving, nor as tolerable as in face recognition. Hence, we can choose C_a = 0.7, C_b = 0.1, C_c = 0.2, C_d = 0.3, C_e = 0.1, and C_f = 0.1.

The Cost of Misclassified Clean Images. As of today, there are no image classification models that can provide a 100 percent accurate result. Table 5.1 shows the accuracy rates of various DNN models without any attacks. ResNet-34 achieves an accuracy rate of 77.47 percent, while VGG16 has a lower accuracy rate of 72.25 percent. On the other hand, DenseNet boasts a higher accuracy rate of 78.69 percent. The percentage of misclassified images is enormous. Therefore, the business models of AI applications should consider these failure cases to assess their risks in case of using any DNN model with a high percentage of misclassification. On the other hand, our

approach can detect a significant fraction of these detection failures and categorize them as forged by adversarial attacks.

As described in the previous section, Dec_b can identify the misclassification of tested samples and be counted as positive to DNN model accuracy. On the other hand, Dec_e, where our approach wrongly identifies it as forged, is counted as negative to the overall accuracy. Application designers can define the costs of these decisions, balancing security and safety with passing tolerance using Algorithm 4. The accuracy of the overall system can be significantly affected, as demonstrated in the following section.

Evaluation Metric. The evaluation technique for our proposed method is similar to those presented in previous works on detection methods [19, 54, 59]. We use the area under the ROC curve (AUC) score in our assessments between clean (Dec_a) and adversarial images (Dec_c), as addressed in Section 5.3. Accuracy (acc.) is another metric used to evaluate our proposed model based on image classification application parameters introduced in Sections 5.2 and 5.3.

5.3 Experimental Results

In this section, we showcase the evaluation and outcomes of our study. First, the settings for the experiments and the utilized environment are explained. Then, the configurations for the adversarial attacks we deploy to target the various deep learning models are outlined. Lastly, we present and compare the main results according to each proposed approach in Sections 5.2 and 5.2.

Experimental Setup. To evaluate the robustness and effectiveness of the proposed scheme, we run our training, evaluation, and attacks using an NVIDIA GeForce RTX 3090 GPU. We use the Sklearn [73] open-source Python library for the classical ML random forest model. On the other hand, we use PyTorch-lightning [29] for DNN models. Finally, we use Torchattacks [47] to run the adversarial attacks.

Adversarial Attack Configuration. The attacker knows that the targeted image classification sys-

tem uses ResNet-34 to train the image classification model. He/she also knows the data being used for that training, i.e., the CIFAR-100 training set. The attacker uses a test set of the same dataset and state-of-art adversarial attack algorithms: FGSM [33], Deepfool [64], CW [15], and PGD [60]. The parameters of each type of AML are listed in Table 5.2 and defined in the next section.

Table 5.2: Experiment settings.

Targeted Model	**Dataset**	**Adversarial Attacks**	**Parameters**	**Attack Success Ratio (%)**
ResNet-34	CIFAR-100	FGSM [33]	ϵ = 0.007	65.75
		Deepfool [64]	s = 50, overshoot = 0.02	99.92
		CW [15]	c = 1.0, κ = 0, s = 50, lr = 0.01	98.64
		PGD [60]	ϵ = 0.03, α = 0.004, s = 40	98.83

In the FGSM trial, we set the ϵ parameter, which is a hyperparameter determining the size of the perturbations introduced to the input data, to 0.007. The value of ϵ is a tradeoff between the adversarial attack strength and the perturbation perceptibility. Raising this value could increase the exploit success rate; however, it might show apparent noise on the targeted image that could be revealed to human perception. We set the default value to 0.007 because the added perturbations are not easily perceived by human eyes. The FGSM attack success accuracy based on the selected ϵ on the CIFAR-100 test set is 65.75%.

To execute the Deepfool attack, we limit the attack iterations to 50 steps before stopping. During each iteration, the attack calculates the direction of the closest decision boundary to the original input data point in order to determine the minimum perturbation required to deceive the targeted DNN model. The overshoot parameter is set to 0.02, which multiplies the computed perturbation vector and adds it to the input image. With these settings, the attack success accuracy reaches 99.92%.

To ensure a successful attack by the CW method, we utilize the CW attack parameters listed in

Table 5.2: $c = 1$, κ = 0, steps $s = 50$, and lr = 0.01. The 'c' hyperparameter determines the magnitude of the perturbation, while the margin parameter (κ) determines the confidence gap between the predicted and target classes. The steps (s) parameter represents the number of iterations required for the attack to succeed or end. Lastly, the learning rate (lr) controls the optimization iteration steps. With these adjustments, we achieve an attack success rate of 98.64%.

The PGD attack is adjusted with the following parameters: ϵ = 0.03, alpha α = 0.004, and steps = 40. ϵ; steps were explained in previous attacks, while alpha functions similar to the learning rate determine the size of each optimization step. This attack has a success rate of 98.83%.

Main Results. Table 5.3 summarizes the AUC scores of four adversarial attack detectors with our proposed method from Section 5.2 using features from all the DNN penultimate layers. For comparison, we compare our proposed method with four other popular adversarial detection methods: DkNN [70], LID [59], Mahalanibis [54], and NNIF [19].

Table 5.3: AUC score of adversarial detection methods.

Detectors	**AUC Score**			
	FGSM [33]	Deepfool [64]	CW [15]	PGD [60]
DkNN [70]	93.65	76.71	93.77	73.78
LID [59]	80.68	52.25	67.84	72.25
Mahalanibis [54]	83.90	62.05	71.60	72.46
NNIF [19]	87.23	84.20	94.58	83.09
Top_1	86.62	**97.57**	**98.21**	**96.49**
Top_22	**94.17**	74.17	83.50	86.04

Overall, our proposed Top_1 threshold surpasses other methods in most attacks, as indicated in bold, while the LID method is the least effective in detecting attacks. The best FGSM attack detec-

tion corresponds to our proposed Top_22 method. Additionally, the NNIF model is the second-best detector approach to resist all attacks. The AUC scores at FGSM show a roughly 10 percent gap between the detectors. In contrast, in Deepfool, the gap is much more pronounced, with LID scoring 52.25 and our proposed Top_1 scoring 97.57.

The AUC score comparisons for different adversarial detector models on various attacks are shown in Figure 5.3. The x-axis represents the four adversarial attacks, while the y-axis describes the AUC score, ranging from 0 to 100. Each color on the graph represents one defense method, as represented in the top-right legend, namely DkNN, LID, Mahalanibis, NNIF, proposed[Top_1], and proposed[Top_22] represented by gray, navy, light green, light pink, light blue, and light brown, respectively. In the FGSM attack, DkNN and the proposed[Top_22] method were the most effective defense mechanisms, while the others showed slight differences, with a score of around 80. The Deepfool bars show significant improvement in detection methods, but some methods have noticeable weaknesses. For the CW attack detection, DkNN, NNIF, and the proposed[Top_1] perform well, while the others score an average of 70. Finally, the PGD attack is proven to be powerful against DkNN, LID, and Mahalanibis, with a semi-matching AUC score of 72, while the remaining methods show significant improvement, with the proposed[Top_1] method scoring the highest, with a score of 96.

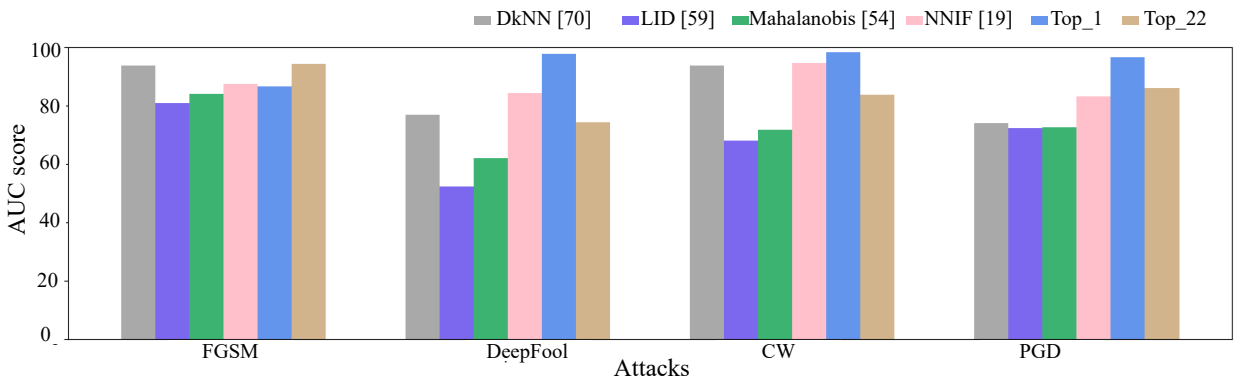


Figure 5.3: AUC score comparison for adversarial attack detectors. The x-axis represents the detector methods. The y-axis represents the AUC score of adversarial detectors. Each color demonstrates one of the detectors, as listed in the top-right legend.

For our proposed system, there is an inherent tradeoff between higher accuracy in detecting adversarial attacks and higher classification accuracy for clean data inputs, as illustrated in Table 5.4. We assign varying weights to each application depending on the potential risks we might face in the event of overlooking a successful attack and depending on our preferred accuracy in classifying normal clean inputs. Our adaptive optimization algorithm (Equation (5.1)) determines that the optimal settings for the RF Top_x probability index should be as follows. In all types of attacks, autonomous driving takes the Top_1 due to the potential for severe accidents if adversarial or misclassified samples are not detected. In health care, FGSM takes Top_5, and the remaining attacks all take Top_1. A face recognition application selects Top_3 for FGSM and Top_1 for the rest. Finally, detecting inappropriate content on a system selects Top_14 for the FGSM attack and Top_1 for the other attacks.

Table 5.4: AUC score comparison based on different applications preferences.

Applications	**Accuracy based on best Top_n selection from formula (1)**							
	FGSM [33]	acc.	DeepFool [64]	acc.	CW [15]	acc.	PGD [60]	acc.
Self driving	Top_1	81.14	Top_1	89.84	Top_1	89.68	Top_1	90.04
Healthcare	Top_5	77.66	Top_1	89.84	Top_1	89.68	Top_1	90.04
Face Recognition	Top_3	79.06	Top_1	89.84	Top_1	89.68	Top_1	90.04
Improper content	Top_14	70.14	Top_1	89.84	Top_1	89.68	Top_1	90.04

As discussed previously detecting misclassification samples could improve the accuracy of the ResNet-34 model. To demonstrate this, we conduct an FGSM attack experiment using the same applications and weights as in Table 5.4. We present the results in Table 5.5. First, we calculate the accuracy without the misclassification samples using Equation (5.1). Then, we calculate the accuracy again after including misclassification samples ($C_b \cdot N_b$ and $C_e \cdot N_e$), as displayed in Table 5.5. Our approach effectively enhances the AML detection accuracy on the ResNet-34 model, which was initially predicted with 74.47 percent accuracy.

Table 5.5: AML detection accuracy comparison before and after including misclassification samples.

Applications	**w/o misclassification (%)**	**with misclassification (%)**
Autonomous driving	62.81	81.14
Healthcare	63.20	77.66
Face Recognition	61.66	79.60
Inappropriate content	60.08	70.14

5.4 Discussion

This section links our proposed ideas with the results and provides a more insightful summary and discussion. We begin by justifying the models and the obtained results. Following this, we elaborate on our model analysis, utilizing a high-accuracy DNN model. Next, we present the challenges associated with this research. Lastly, we introduce our future plans for this project.

Justifications. Although the RF is a sufficient model in regression [56] and multiclassification applications [8], it is not commonly used for image classifications because images have a large number of pixels, resulting in high-dimensional feature spaces. In addition, image processing is computationally expensive and time-consuming during training. However, we decided to use RF as a secondary model for two reasons. Firstly, other models such as support vector machine (SVM) [20] are not efficient in multiclassifications and are computationally expensive. Secondly, we want to showcase the usefulness of having two different architectural models to overcome adversarial attacks. Studies such as [19, 70] have used traditional ML algorithms to create AML detectors. They adapted k-NN in generating their adversarial detectors by adding a k-NN model between DNN layers during training to extract new features that can be analyzed and used to recognize clean and noisy samples versus adversarial ones. However, in addition to k-NN's extreme complexity and high computational performance, these studies found that different types of attacks have varying resistances depending on the effectiveness of the attack in generating perturbations to fool the model.

For example, Table 5.3 shows fluctuations in AUC scores in resisting each attack by every detection algorithm, such as NNIF for FGSM, Deepfool, CW, and PGD, with scores of 87.23, 84.20, 94.58, and 83.09, respectively. In contrast, our proposed system with the Top_1 setting is consistently effective, regardless of FGSM outcomes, as it has a large number of clear samples that are not affected by the attack at α 0.007; further clarification is provided later in this section. Therefore, a new technique of attack that relies on backpropagation could harden the defense algorithms, as illustrated in [19], when a detector trained on an FGSM attack is only tested on unseen attacks

such as Deepfool. These findings indicate a decrease in performance when testing for unseen attacks compared to seen attacks. Our proposed system, however, is tested on all adversarial attacks without attack pattern evaluation nor DNN model changing and presents a generalization across different attacks.

Additionally, the results of the FGSM attack in Table 5.4 show reasonable changes with Top_k based on an application's weight parameters. This change from Top_1 in autonomous driving to Top_14 in detecting inappropriate content is normal when we increase the cost of C_f. In this situation, Equation (5.1) significantly increases the security sensor to minimize the success rate of adversarial attacks. In contrast, the equation reduces the model sensitivity when preventing inappropriate content because the risk of successful attacks is not so serious. This equation provides freedom to the application developer to choose the best and most optimal defense setup.

Unlike the other attacks, Table 5.4 shows that we consistently use Top_1 for every test of Deepfool, CW, and PGD attacks because of a couple of reasons. First, the variation in the success rate of these attacks, as shown in Table 5.2, is based on the attack strategy and the strength to fool the model. For instance, PGD is developed from an FGSM attack, where PGD has a selected number of iterations to break the model, while FGSM applies one-time manipulation based on the ϵ value. Additionally, the success ratio is exemplary or unrealistic. In real-world attacks, attackers have no access to information about the ML models, the data used for training, the integrated security level, etc. The regular success rate should be much less than that in the examples presented in the table. An FGSM attack is an example of the successful usage of our proposed adaptive design theory; otherwise, the system's adaptivity is useless if the applied systems are extremely exposed to adversarial attacks.

Model Scalability. Our experimental construction is based on the ResNet-34 model, which has an accuracy of 77.47 percent. We select this model for our demonstration to match previous experimental setups and compare our output enhancement. To elaborate and to present our approach efficiently, we train the ViT-B_16 model on the CIFAR-100 dataset and achieve 92.58 percent ac-

curacy. Then, we attack the model with an FGSM attack. Our defense shows validity in detecting adversarial attacks, for which we obtain 78.90 with Top_1 and 89.49 with top_40 based on the AUC score. Comparing ResNet-34 with ViT-B_16, our approach selects smaller k on ResNet-34 for the following reasons. First, the accuracy of ViT-B_16 before the attack is higher by 14 percent. Next, the overall accuracy of random forest is low compared to DNN models, which makes k changes regarding the primary model accuracy. Finally, the classical model accuracy is careless since the DNN model is the main model the application relies on for classification, and we adapt the idea of k to overcome this challenge.

Challenges. Challenges are found throughout this study. Foremost, the RF model is designed professionally for structural and tabular applications such as stock market price predictions that use a specific number of vectors; simple image classification; or recognition tasks, such as satellite imagery object detection. However, the RF model capability can be limited when faced with large-scale training involving a significant number of classes, such as when using the ImageNet dataset [24] with its 1000 classes and 1.2 million images. To tackle these challenges, a DNN model was developed. It excels in extracting features from high-dimensional vector spaces and large datasets while requiring less time for training. We use this RF model as a prototype for our analysis, and we highlight that a potential robust defense mechanism exists if we can adopt a different architecture.

CHAPTER 6: IMAGE AUTHENTICATION IN EXTRA-MILITARY: CHALLENGES, OPPORTUNITIES, AND CASE SCENARIO

In this chapter, we discuss the different types of big data used in the military. We provide a flowchart for fusing data from various sources, including public ones like social media platforms. The flowchart helps gather, prepare, authenticate, and provide the data for decision-making to expedite military operations. This study examines the challenges involved in obtaining data from various sources and its security, including authentication of images using our approach of image authentication. We also demonstrate the impact of our approach on improving decision accuracy. Additionally, we discuss other security challenges and suggest solutions to them. [1]

6.1 Big Data type in the Military Domain

To fully comprehend the potential of big data in the military context, we first need to identify the data categories that are directly and indirectly related to the field. Drawing from the ideas presented in [81], which focuses on transportation data, we have introduced the concept of Multiple Data Sources (MDS). This concept is then used to describe the military scenario and support decisions based on different data sources. The MDS is classified into two main groups: IMD and EMD, as illustrated in Figure 6.1. This taxonomy is a basis for discussing how the EMD can support strategic military decisions, which was only made possible due to the significant data phenomena. Furthermore, we delve into the different aspects that influence the use of EMD, such as data privacy/security, integrity, data acquisition, fusion, and learning approaches.

[1] The contents of this Chapter are based on our publication that is submitted to IEEE ComMag 2023

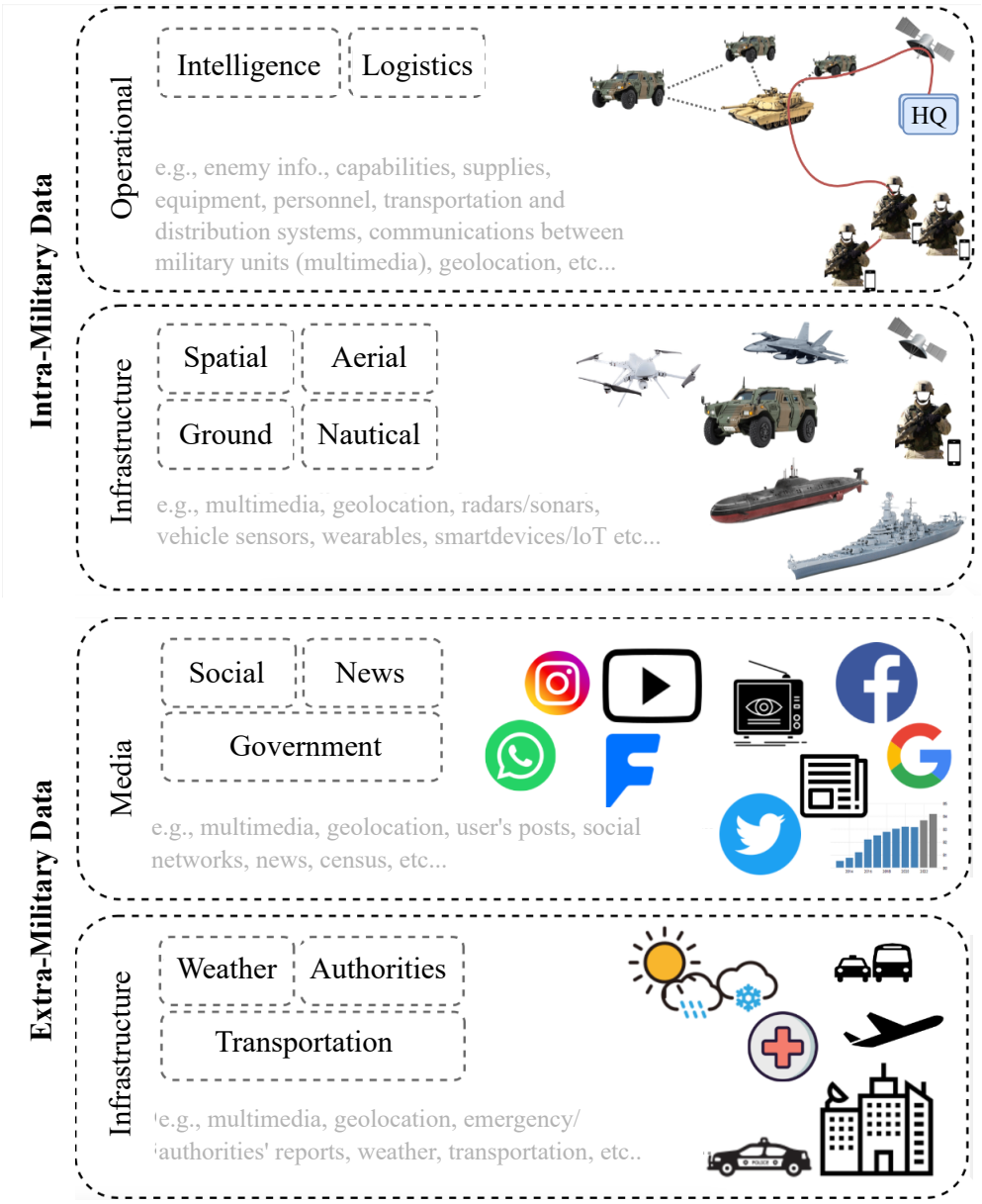


Figure 6.1: Military Data Space.

6.1.1 Intra-Military Data

The IMD data used by the military comes from two main sources: infrastructure with real/virtual sensors and operational data, such as intelligence and logistics. The infrastructure of the military involves the use of various electronic systems to detect and track objects on land, in the air, or in

the water. These systems include:

1. data collected by sensors such as radar, sonar, cameras, and other electronic systems that can detect and track objects in the air, on land, or in the water,
2. vehicular sensors that provide the status of the military units and
3. wearable/smart and Internet of Things (IoT) devices support the infantry in the field with GPS position, maps, health measurements, live cameras (high resolution, infrared), etc

The data collected can be used for a variety of purposes, such as identifying potential threats, monitoring activity, targeting enemy forces, and keeping track of infantry conditions. The Integrated Master Data (IMD) does not only rely on raw data from sensors but also combines information obtained from various sources to create a more dependable and unified operational view as required by the Joint All-Domain Command and Control (JADC2) and Common Operational Picture (COP) systems. The intelligence data can also help military forces understand the enemy's capabilities and intentions, identify possible dangers, and plan their operations accordingly. Furthermore, the logistics data provides essential information on supplies, equipment, and personnel, including transportation schedules, inventory levels, and maintenance records. This information is crucial to ensure that the military forces have the necessary resources to carry out their missions effectively.

6.1.2 Extra-Military Data

The EMD stands for the subset of data provided by real and virtual sensors, either individually or fused. This data can describe the environment around the military operation, not just geographically but in other ways as well. Two primary sources of data can support military operations: the media (such as social media, news, and government reports) and the infrastructure (such as transportation systems, weather, and authorities). These sources generate a large amount of highly variable data, ranging from user-generated content like photos and feelings related to real events

such as accidents and terrorism, to traffic and weather conditions, as well as people's and drivers' behavior.

The rapid urbanization that has occurred in recent years has put a strain on Information and Communications Technology (ICT) to enhance the quality of life for people living in cities. This has resulted in the emergence of the concept of Smart Cities, which proposes solutions to the problems that have arisen from urbanization, such as issues with mobility, safety, and health. The increasing number of sensors and mobile devices in urban areas (known as IoT) has resulted in the creation of large amounts of data, also known as Big Data. Communication technologies have made it possible to use this data through the cloud to design intelligent and efficient systems that can cater to people's mobility, health, safety, and economic demands.

Based on the potential advantages of smart cities, we propose that the military can leverage the knowledge and data offered by civilian systems in urban areas across all aspects of their operations, including the political and ground level. We further suggest potential scenarios where such synergy could be useful.

Smart cities are equipped with a variety of sensors, including inductive loop traffic detectors, cameras, radars, traffic lights, weather sensors, and authorities reports about emergencies and disasters. These sensors provide information about the state of vehicles, traffic conditions, weather, and drivers' behavior. Inductive loops are a type of sensor that can detect when a vehicle passes a certain point and its speed. They can also be used to classify types of vehicles based on their unique signatures. Cameras and radars can detect a vehicle's speed and type. Additionally, cameras use advanced recognition algorithms powered by AI to detect potential threats such as terrorism, incidents, and disasters and notify authorities accordingly. The infrastructure can also provide information about groups of people and has been used in surveillance systems to identify potential threats at different levels of detail. Therefore, the civilian infrastructure can provide contextual information and has the potential to support tactical and strategic military domains.

The concept of data fusion has been discussed and explored in the literature. With access to various

types of data, it aims to enhance data quality and coverage by adding different perceptions and descriptions of an event through multiple data providers. For instance, in the event of a disaster in an urban area where civilians are possibly injured, the fusion of data from various sensors available in a city, such as transportation, weather, cameras, and health systems, can assist military operations and smart civilian applications. This can be done by providing valuable information such as identifying non-damaged roads and fast paths for rescue operations, forecasting weather for the rescue mission, reserving specialized medical treatment nearby, and monitoring the causes of the event to track possible terrorist actions.

Although infrastructure sharing can provide reliable data, it may not always cover all areas to describe events accurately. Therefore, media sources such as social media, news outlets, and government reports can be used to understand how the locals behave and identify socioeconomic factors that may contribute to criminal activity or prioritize assistance in a particular region.

Numerous initiatives have been launched in recent years, partly in response to Russia's war and the rise of extremist movements against democracy in countries such as the U.S. and Brazil. An example of such initiative is the non-profit project ACLED (Armed Conflict Location & Event Data - `https://acleddata.com/`), which provides real-time, global data on political violence and protest events. ACLED analyzes and publishes high-quality, disaggregated data on conflict and protest activity in over 100 countries. Researchers, policymakers, and journalists widely use their data to track and analyze events.

A notable example of EMD is the DATTALION database (`https://dattalion.com`). It is the largest open-source collection of photo and video footage from the war between Russia and Ukraine. The primary aim of this database is to counter the spread of misinformation by the Russian government. The photos in this dataset are categorized into three groups: (1) Official, collected from verified sources such as government officials, city mayors, regional heads, and government organizations. (2) Trusted, coming from sources that are considered indirectly verified, like reputable news outlets, reliable Telegram channels, and official pages of journalists. (3) Not-

Verified, gathered from sources that do not have their own content verification procedures. These could be Telegram channels, personal pages and groups on social media, Twitter accounts, and other social network platforms.

The United Nations Development Programme (UNDP) utilizes machine learning algorithms and big data to identify infrastructure damaged by war in eastern Ukraine. The algorithms are trained by a semantic damage detector, which uses satellite imagery and ground-based photos to analyze and identify potential damage to buildings, roads, and bridges in new images. The purpose of this project is to assist local authorities and humanitarian organizations in prioritizing their rebuilding efforts and improving the accuracy and speed of their assessments. The project has successfully identified and mapped damaged infrastructure, and it is expected that this approach can be applied in other areas affected by conflicts.

EMD encompasses various data categories, including social media data that local users share. This data can be utilized to obtain information that cannot be captured by other sensors. For instance, it can be used to locate groups of people, who may be injured or hiding and require rescue. Stationary sensors on buildings and surveillance cameras can also be used to track humans and identify their location. Combining social media data with other sources can help detect enemies and plan effective maneuvers against them. Hence, all these sensors can provide a more comprehensive view of the disaster/incident situation, which can aid military operations.

Transportation-related sensor data, such as traffic surveillance cameras in smart cities, can also support emergency rescue operations and military logistics. By analyzing traffic data, congestion or blockages due to incidents like accidents or protests can be detected. This leads to better route planning and traffic management, such as controlling traffic lights and digital signs, during military operations. Additionally, fusing all the collected information can improve the overall understanding of the event and enhance operational planning and management in urban areas.

6.2 Use Cases

This section will discuss two use cases that utilize EMD to enhance the available information. We achieve this through a data fusion framework that combines different data sources and maintains the integrity of images shared via social media. This approach supports strategic decision-making based on COP and systems like JADC2.

6.2.1 Data Fusion

Methodology. In order to support military decisions, it is important to fuse big data in a spatiotemporal manner. To achieve this, we have designed the Fusion Multiple Data Sources (FMDS) framework [121]. This framework fuses various types of data sources and performs functions such as collecting, preparing, and processing data to generate enriched information. We have implemented FMDS to demonstrate the enrichment of spatiotemporal data using transportation system data. However, it is not limited to this subject and can be extended to different types of data. The ultimate goal is to improve data quality, enhance Command and Control (C2) systems, and support the COP/JADC2. In the following, we will briefly describe the main functionalities of FMDS, as illustrated in Fig. 6.2. We will also discuss the benefits of fusing big data by presenting numerical results.

- **Data Acquisition** FMDS collects data from various sources and stores it in standardized CSV files using a set of configured parameters such as region and request frequency (Fig. 6.2 (1)).

- **Data Preparation** The input dataset undergoes a standardization process that converts different feature names and types into a uniform representation, as shown in Fig. 6.2 (2). This process includes various data mappings that can be configured within the framework setup to generate uniform data types. For instance, it involves mapping descriptive values into a numerical representation or reducing data granularity. The map-matching process is also

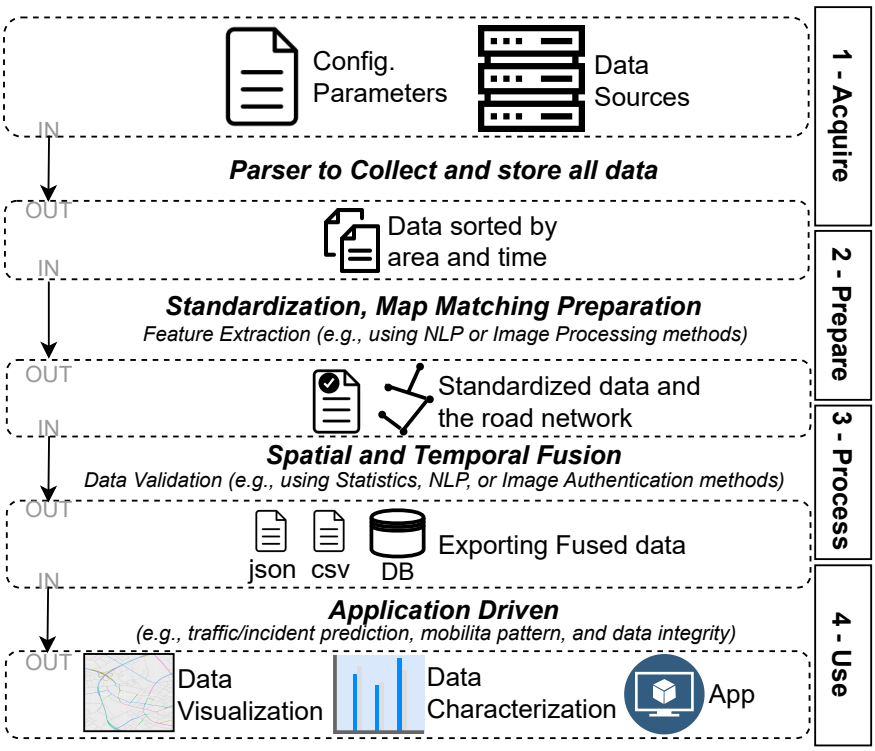


Figure 6.2: Workflow of the data fusion framework.

prepared, which involves creating all input files required, including a Shapefile (SHP) from OpenStreetMap (OSM) based on a defined geo-location. Additionally, the collected information is transformed into a usable representation for the map-matching procedure. In certain cases, the framework utilizes feature extraction methods such as Natural Language Processing (NLP) or image processing to extract the information from a given data type.

- **Data Processing** In Fig. 6.2 (3), we see the data processing that involves three sub-tasks, namely, temporal data fusion, spatial data fusion, and data export. To ensure data integrity, we use various data processing methods, such as validation of incident reports or image authentication (discussed in Section 6.2.2), to filter out non-trusted or biased information before the fusion process.

  For temporal fusion, we group data within an arbitrary time window, such as 10 minutes, hourly, or daily, to allow comparison of different data types. For spatial fusion, we use map-

matching, a process that aligns GPS points under a defined degree of accuracy based on an underlying road network. This is necessary due to the varying precision of GPS reports resulting from different data sources. Map-matching provides a balanced accuracy level on the same road network.

The enriched and updated geo-information is then added to the input dataset. The data fusion process combines the fused input dataset with the map-matching output, using the primary key and different grouping methods to optimize the process.

- **Data Usage** Enriched data can be exported in various formats, providing ample opportunities for both military and civilian applications. This data helps to enhance the information available for a given urban area, enabling better decision-making in the context of smart cities and military operations. The FMDS outputs support spatiotemporal analysis through the creation of statistics and visualizations such as heat maps, density plots, and bar graphs, which characterize the available information under various spatial and temporal aspects.

Results. In order to showcase the advantages of data fusion, we will present the results of a previous evaluation conducted on the FMDS framework. The experiment gathered four types of civilian transportation data from seven different providers over a period of nine months. The framework was able to increase the data coverage by 173.

Moreover, the combined data was utilized to conduct a thorough examination of the gathered information. Such data analysis enables the detection of specific issues and can aid in devising strategies to enhance transportation networks. We presented a range of statistics and visualizations to contribute to the analysis of spatiotemporal data.

We presented two data applications using the fused dataset, demonstrating data fusion's support of such applications. Our evaluation validated this hypothesis, achieving positive results for both traffic estimation and incident classification.

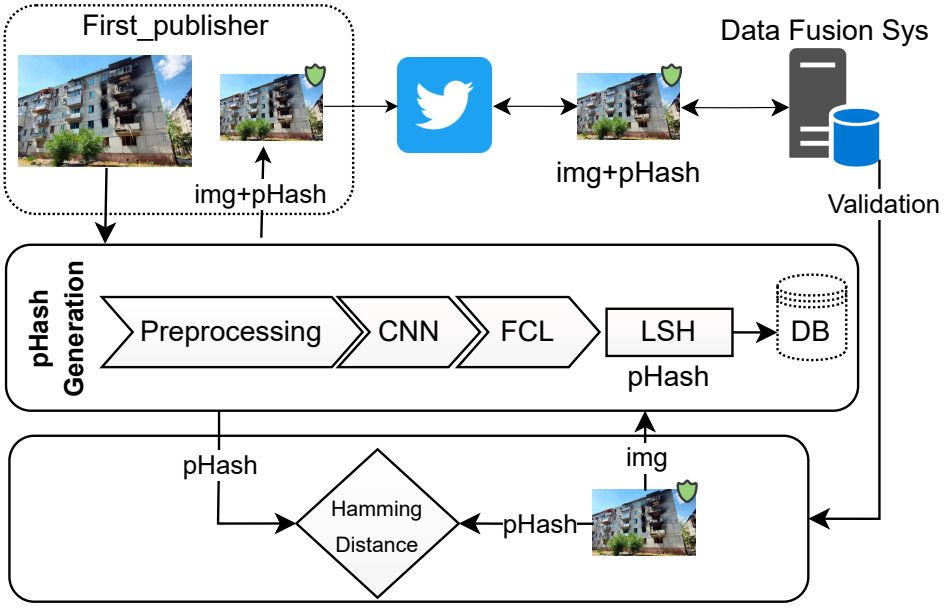


Figure 6.3: Image-Fact-Checker (IFC).

6.2.2 Image Authentication

Methodology. In our previous work [3], we presented an image authentication system that can be used in real-world applications. This system utilizes the Twitter and Facebook platforms to ensure the integrity of images. To generate a sensitive Perceptual Hashing (pHash) for detecting any content manipulation, we used the SMPI images dataset [6], which was collected from these platforms. Our model employed a Convolutional Neural Network (CNN) and Fully Connected Layers (FCC) to extract image features, and Locality Sensitive Hashing (LSH) to construct the final hash. For better accuracy, we also implemented contrastive loss, which maximizes the differences between original and manipulated images, and minimizes the similarity between each category and its augmentations. The final output of the model is a fixed-length vector representation for an image, which has a length of 1024 bits.

Ensuring the integrity of images is crucial to support both urban military operations and civilian

systems. In the absence of verification, the combination of multiple sources can be vulnerable to manipulation. In order to supplement our data fusion framework, we have developed the Image-Fact-Checker (IFC), which can detect any fraudulent images and thereby ensure the reliability of the data, as depicted in Fig. 6.3. IFC functions as an authentication system that is authorized by relevant authorities to combat the spread of misinformation.

In order to maintain the authenticity of images, we believe that the first person to publish an image (such as a social media platform) should play a collaborative role in reducing the spread of false information. This can be achieved by uploading the image to the system for the first time. The system will then generate a new version of the image that includes a logo or an icon indicating that the image has been verified before publication. Users on social media platforms can validate the image through the IFC system. Additionally, IFC provides a pHash string representation of the image that can be included in the post's description or on other Internet websites. The data fusion system in Figure 6.3 acts as an end user of IFC. It validates crawled images from social media platforms to determine if IFC can trust them before applying spatiotemporal fusion to produce enriched data outputs.

Results. Utilizing the IFC system increases the reliability of the data provided. Additionally, we can ensure the authenticity of each posted image from a single source to detect any manipulation. In the event of a war, such as the ongoing conflict between Ukraine and Russia, imagine an attack on a city where civilians reside. Towers and buildings are set ablaze or destroyed, and some individuals are injured or manage to escape. Some may make an emergency call, while others may use their phone cameras to capture images of the attack. As we have seen since the war began on February 24, 2022, numerous images of the conflict have been posted on social media and collected by DATTALION. However, due to trust issues with the initial publishers, the utilization of these images is somewhat limited.

Now, if we pass the data collected by DATTALION through IFC and have each image verified and sent to the relevant rescue organizations, the United Nations, NATO, or other organizations

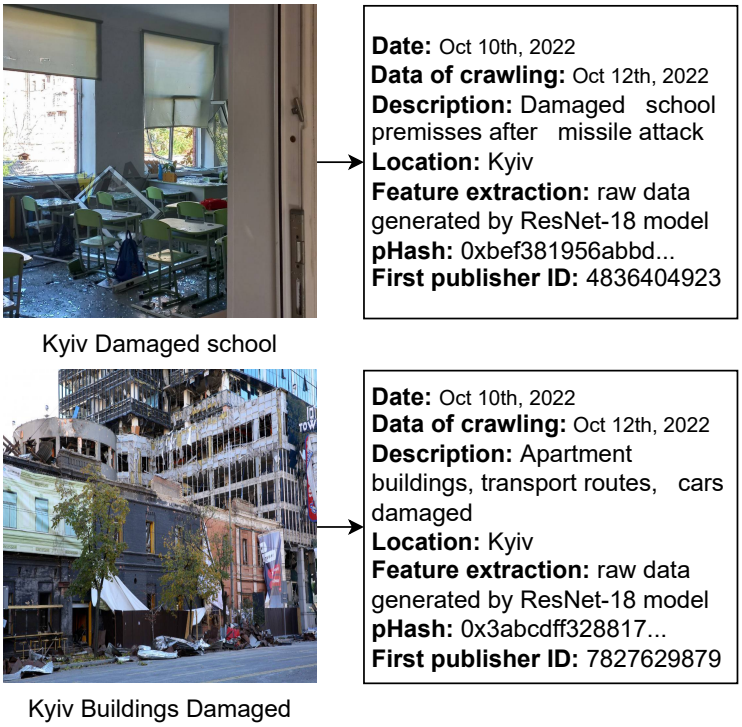


Figure 6.4: Extracted image details with IFC.

Source: Dattalion dataset

affiliated with the war, we can respond to each attack more quickly and reasonably. Furthermore, the verified images can also be used as valid evidence against the Russian regime.

In Fig. 6.4 (left), you can see two images that were shared on social media by regular users. These images were collected from a data source called DATTALION, which has many images that cannot be trusted due to their unknown sources. However, if we put these images through the IFC mechanism, we can increase the reliability of the dataset. The IFC mechanism adds a pHash to each image along with its related information like the image description, extracted feature, location, event date, date of crawling, and first publisher Twitter ID (as shown in Fig. 6.4 on the right). Afterward, these images are stored in the IFC database and can be used for duplicate detection, verification, and other applications as needed.

6.3 Challenges and Opportunities of Big Military Data

This section delves into the possibilities and challenges associated with the application of big data in the military scenario, specifically EMD and IMD. It discusses crucial aspects like data fusion, security, privacy, integrity, and the utilization of AI while proposing directions for the development of the next generation of intelligent military applications.

6.3.1 Data Fusion

The process of merging the different types of data that provide information to the MDS is still challenging and needs further processing and treatment. In order to make use of the additional information obtained from various data sources, it needs to be standardized into a consistent format, which permits spatiotemporal fusion. Standardization involves several data mapping techniques, changes in granularity, and conversion of features and time formats. When working with images and text from the media, we also require feature extraction methods such as NLP and image processing to extract useful information from these data types.

6.3.2 Data Security, Privacy, and Integrity

Data Security and Privacy. Ensuring security and privacy is a top priority in our system design, as it involves gathering and storing information in a database. Security measures are put in place to prevent unauthorized access and modification of data within the system. As end-users need to establish remote connections to access the data fusion platform, they are exposed to various threats that can cause disruptions, alterations, or even destruction of the system's digital services. One such common threat is a Man-in-the-Middle (MITM) attack, which enables an attacker to hijack the connection between the user and the server for nefarious purposes such as manipulation or eavesdropping.

Protecting personal information is important for individuals' privacy. It includes sensitive data

like names, addresses, and social security numbers. While gathering data from open sources like Extra-Military Data can improve system security, it also raises privacy risks for regular users. They could become vulnerable to attacks, which might reveal or steal their data. According to a report by IBM Security and Ponemon Institute[2], cyberattack costs have risen by 2.6% in 2022 compared to the previous year. The average cost of a data breach globally is 3.35 million US dollars. This shows that most organizations have experienced cyberattacks, with 83

Ensuring the security of our digital systems is of utmost importance. We can achieve this by implementing various techniques such as PKI security infrastructure, data and user information encryption, and incorporating firewalls and intrusion detection tools. By integrating these security measures into our system, we can significantly reduce the system's vulnerabilities and enhance the safety of both data and users.

Data Integrity. Ensuring data integrity in the Fusion MDS is crucial to prevent manipulated data from impacting both civilians and military operations and to maintain trust in data providers. Obtaining data from social media platforms like Twitter and Facebook increases the likelihood of receiving misinformation. Fake news is intentionally created to persuade people to adopt erroneous opinions, and propagandists often use it to spread false information and exert political influence. The war statistics in Ukraine is a prime example of the widespread dissemination of disinformation from the very beginning of the conflict, leading social media platforms like Twitter to update their policies and provide reliable data to their users [61]. Twitter reported that within the first month of the war, it received around 38.6 billion impressions of real-time tweets and identified over 900,000 unique tweets containing links to Russian state-affiliated media. This put Twitter in a difficult position where it had to restrict some services for the warring parties.

Images on various platforms can be used to convey complex ideas and events quickly. For instance, a single photo of a threat can immediately communicate the need for emergency rescue operations, allowing quick action to save lives or change transportation pathways. Moreover, images can add

[2]https://www.ibm.com/downloads/cas/3R8N1DZJ

emotional context to a news story, making it more relatable to readers. However, manipulated images often spread faster and receive more interactions than real ones. A study of Twitter images during Hurricane Sandy [35] revealed that approximately 90% of retweets were from tweets with fake photos.

The extensive media coverage of global crises such as the Ukraine war increases the risk of spreading misinformation. To counter this, public fact-checking websites like `snope.com` and `norc.org` investigate critical news worldwide using human resources. However, in the case of real-time events during wars, it can take a long time to manually verify sources when unknown individuals on social media platforms share images. Hence, there is a pressing need and an opportunity for automated systems to provide image authentication and detect misinformation.

6.3.3 Artificial Intelligence (AI)

The rapid progress of AI applications owes much to big data and computing power availability. The quality of AI depends on the data fed to Machine Learning (ML) models during training. Generally, the more data we have, the better we can train the model to avoid overfitting. However, the limited availability of public data in the military domain poses a challenge to studying past and future developments. Although we expect that the revolution in autopilot, drones, intelligent missiles, security surveillance cameras, etc., relies on big data owned by military agencies, the limited availability of this data can be explained by the need to protect individual's privacy and security and prevent misuse by radicalized, violent, or enemy groups.

Apart from the limited availability of big data, there are other challenges to consider for safety and security in AI, such as the potential failure of AI functionality due to hacking incidents. The architecture of AI systems, such as DNN used in computer vision applications, must be more secure, posing a significant challenge. Although this approach pushes computer vision applications forward, the complexity and sophistication of DNNs make it increasingly challenging to ensure their security against adversarial attacks. To date, no single method or approach has been proven

to be completely secure against such attacks.

An adversarial attack is a type of cybersecurity threat that specifically targets AI systems. Attackers use deceptive data to fool ML models, which can misclassify manipulated data that human eyes cannot recognize. Adversarial attacks are categorized into three main groups based on the attacker's knowledge of the targeted AI system: white-box attack, grey-box attack, and black-box attack. Techniques such as the Fast Gradient Sign Method (FGSM) for computer vision and semantic attacks for NLP are used to identify adversarial attacks in various AI applications.

Several defense mechanisms have been introduced recently to tackle one or more adversarial attacks. The authors of a recent study categorized adversarial defense mechanisms in computer vision into three groups: a defense that targets the model by modifying the model itself for robustness against adversarial attacks defense of modifying the input sample for perturbation removal defense of adding external modules for adversarial sample detection In addition, Federated Learning (FL) has emerged as a technique for training ML models, where the data cannot be exposed, ensuring data security and privacy. Although FL cannot be considered a defense technique against adversarial attacks, it can hide sensitive data and expose only part of the model or parameters, making it a valuable technique for military applications.

6.3.4 Networking

Although the primary focus of this work is centered on the data perspective and the importance of utilizing reliable data from various sources to support military operations, it is equally important to recognize the significance of networking in delivering data and services effectively. Communication systems that are dependable and efficient are crucial in both military and civilian operations, particularly in remote and challenging environments. Wireless communication is especially important in network-centric military operations. These networks incorporate various technologies such as High Frequency, Very High Frequency, Ultra High Frequency, Satellite Communications, Wi-Fi, and LTE 4-5G. While some of these technologies excel in long-range coverage, they have

limited bandwidth, high latency, and are susceptible to disruptions. Conversely, others prioritize reliability with a shorter range, greater bandwidth, and lower latency.

Information-Centric Networking and Software-defined Networking are crucial network paradigms that enhance data dissemination and network orchestration [53]. These paradigms emphasize the importance of data fusion in disseminating data with limited network resources. In the military network infrastructure, especially in tactical situations, there can be issues such as limited network resources, security concerns, high delays, and low bandwidth when disseminating data. To overcome these limitations, the military is considering using civilian networks such as 5G as supplementary infrastructure to gather extra-military data and share non-classified information.

The recent conflict between Russia and Ukraine has highlighted how vulnerable communication networks can be. Russia's attacks on infrastructure caused internet outages, which can be detrimental during wartime. SpaceX's Starlink has come up with a solution to this problem with its satellite internet constellations. This technology has demonstrated the benefits of using civilian network infrastructure during times of war. Although it shows promise in enhancing internet reliability for both data and emergency communication, it still faces challenges in terms of cybersecurity, coverage, reliability, and cost-effectiveness.

CHAPTER 7: CONCLUSION

7.1 Summary

The rise of social media has led to a rapid increase in disinformation, including fake images that can significantly impact public behavior, attitude, and belief. Detecting manipulated images has become crucial for maintaining public trust, and effective authentication is essential. This dissertation comprehensively evaluates multiple state-of-the-art perceptual hash algorithms for detecting image manipulation on Facebook and Twitter. Perceptual hashing offers advantages in detecting image manipulation while tolerating typical image processing or resolution changes on user-uploaded images. Our studies show differences in image processing between the two platforms and propose new approaches to finding optimal detection thresholds for each perceptual hash algorithm. Additionally, a new and robust perceptual hash authentication approach is presented, utilizing a self-supervised learning framework and contrastive loss, and outperforming state-of-the-art methods. A fake image sample generator is also developed to cover known image attack types.

Our dissertation presents a new method of identifying adversarial attacks that surpasses existing models. We incorporate a classical machine learning model alongside the primary DNN model for image classification. The secondary model is designed differently from the primary model to counter backpropagation-based adversarial attacks. Our proposed detector does not require any changes to the DNN model or learning attack types, and we tested it on the CIFAR-100 dataset. However, we faced challenges during the study, such as the limitations of the RF model for large-scale training with many classes.

In addition, the dissertation demonstrates the benefits of adapting the image authentication system in a real-world environment by reviewing a military application. It introduces a workflow of decision-making based on fusing data from different resources to accelerate accurate decisions during a war situation. The study covers the challenges and, at the same time, suggests reasonable solutions using state-of-the-art methods.

7.2 Future Work

We anticipate advertising our introduced ideas for public and private organizations to fight misinformation in the cyber world. By that, we should continue to update the work according to the technological changes, study different media types, and participate in continuing to suggest more solutions. The following are open areas to be studied and investigated in depth.

Classical ML applicability & scalability. Scalability in classical machine learning can be challenging, as previously discussed. To tackle this issue, first, we suggest using the DNN model's extracted embedding. After that, the embedding feed to the RF model for training. This strategy may increase the accuracy of RF and create a more effective detector. By combining the strengths of both models, we can benefit from DNN's superior feature extraction and RF's outstanding ability to mitigate adversarial attacks.

DCT defense mechanism. Utilizing the DCT approach to extract feature vectors from a CNN model can be a promising method to prevent adversarial attacks in our future work. DCT has been used in image processing and feature extraction for over a decade. Its simplicity and positive results make us optimistic about integrating it as a layer in the entire authentication pipeline. This could be done before generating the pHash and after receiving the feature value from the SimCLR model, the embedding. We expect that DCT may remove any added perturbations caused by adversarial attacks since it compresses inputs to a short representation.

Different threat model. Our proposed system has a high chance of success, even if the attacker is familiar with the random forest method. This is due to the fact that most adversarial attacks in computer vision rely on gradient descent, which is a primary function in DNN models. Unlike DNN, classical machine learning has a different base, such as in RF, and does not utilize the gradient descent technique. However, we still need to study how attackers can potentially bypass our defense measures, including testing if an attacker can deceive our detection methods while being aware of the second approach used to identify adversarial attacks.

NLP Authentication. The idea of image authentication can be applied to text-based news as well. A possible implementation is developing a pHash algorithm that can determine if a piece of text-based news has been altered in terms of its content, tone, mood, and factuality while still allowing for minor changes in wording. This proposal could help in the fight against fake news, which is a significant concern globally. The architecture design can also be used in language understanding and generation using transformer architecture, such as BERT and GPT.

The suggested design involves a pre-processing stage where sentences are categorized as either positive or negative matches based on their similarity. An encoder is then used to generate the embeddings, such as BERT, and contrastive loss algorithms are employed for learning. Finally, the model is represented as a string pHash that can be used to verify the authenticity of the news.

Moreover, it could be generally easier to train a secondary classical ML model for some NLP tasks compared to image processing because of the size of the image vectors. Images are typically represented as 3D arrays with 32x32 pixel values for each dimension, each with a probability of 256 values for each pixel. Compressing or resizing the inputs can lead to a loss of quality and make it harder for the model to learn the details. As a result, good computer vision models usually have embedding sizes between 1024 and 4048. In contrast, some NLP models, like BERT, use embeddings of size 768 for learning. This makes traditional classical models a suitable approach to defend against adversarial attacks.

Hierarchical-Graded Classification. In traditional classification, models treat all classes as unique and independent of each other. They arbitrarily label N classes with the labels of 1, 2, 3,. . . ., N. When conducting object classification, the result is either ”1” (match) or ”0” (mismatch). However, in reality, each class among the predefined N classes has its own physical meaning and certain relationships with all other classes. In any application field, all classes are correlated with each other. This means that traditional classification does not provide a good representation of such correlations.

On the other hand, in Hierarchical-graded classification, the output is not binary with only a value

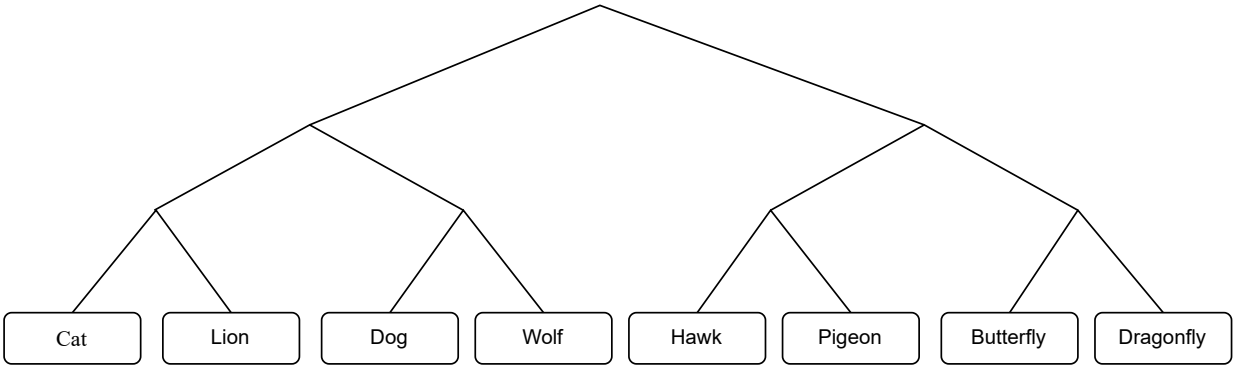


Figure 7.1: Hierarchical-graded classification.

of ”0” or ”1”. Instead, the classification output C will be a real number between ”0” and ”1”. Let us use an example to illustrate the concept. Suppose an animal identification problem has eight classes: Cat, Lion, Dog, Wolf, Hawk, Pigeon, Butterfly, and Dragonfly. In traditional classification, they are randomly ordered and labeled as 1, 2, 3,. . . , 8. In our suggested Hierarchical-graded classification scheme, the application expert first defines the hierarchical structure to represent these eight classes' relationship as in Figure 7.1.

From Figure 7.1, If an object is ”cat,” then the classification value C will be something like, given the classified class by ML as follows:

- ML output = Cat, C = 1, and prediction matches a perfect match.
- ML output = Lion, C = 0.75, and prediction matches at the second level.
- ML output = Dog/Wolf, C = 0.5, and prediction matches match at the first level.
- ML output = Hawk/Pigeon/Dragonfly/Dragonfly, C = 0, and prediction does not match at the first level.

In this way, we can better evaluate ML classification performance, especially when misclassification occurs due to intention attacks, such as adversarial attacks. This is due to the size of the perturbations attackers need to change the correct labels to different labels from different branches.

In addition, This proposed hierarchical-graded classification will be more useful when the application has a large number of classes (such as 50 or 100) and when the misclassification rate is high.

APPENDIX A: PUBLICATIONS COPYRIGHT

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Evaluating Perceptual Hashing Algorithms in Detecting Image Manipulation Over Social Media Platforms
Mohammed Alkhowaiter, Khalid Almubarak, Cliff Zou
2022 IEEE International Conference on Cyber Security and Resilience (CSR)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

You have indicated that you DO wish to have video/audio recordings made of your conference presentation under terms and conditions set forth in "Consent and Release."

CONSENT AND RELEASE

1. In the event the author makes a presentation based upon the Work at a conference hosted or sponsored in whole or in part by the IEEE, the author, in consideration for his/her participation in the conference, hereby grants the IEEE the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive, in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IEEE and live or recorded broadcast of the Presentation during or after the conference.
2. In connection with the permission granted in Section 1, the author hereby grants IEEE the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IEEE from any claim based on right of privacy or publicity.

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Mohammed Alkhowaiter — **Signature**

02-06-2022 — **Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the

IEEE COPYRIGHT AND CONSENT FORM

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IEEE and must accompany any such material in order to be published by the IEEE. Please read the form carefully and keep a copy for your files.

Image Authentication Using Self-Supervised Learning to Detect Manipulation Over Social Network Platforms
Mr. Mohammed Alkhowaiter, Mr. Khalid Almubarak, Mr. Mnassar Alyami, Mr. Abdulmajeed Alghamdi and Dr. Cliff Zou
MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)

COPYRIGHT TRANSFER

The undersigned hereby assigns to The Institute of Electrical and Electronics Engineers, Incorporated (the "IEEE") all rights under copyright that may exist in and to: (a) the Work, including any revised or expanded derivative works submitted to the IEEE by the undersigned based on the Work; and (b) any associated written or multimedia components or other enhancements accompanying the Work.

GENERAL TERMS

1. The undersigned represents that he/she has the power and authority to make and execute this form.
2. The undersigned agrees to indemnify and hold harmless the IEEE from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.
3. The undersigned agrees that publication with IEEE is subject to the policies and procedures of the IEEE PSPB Operations Manual.
4. In the event the above work is not accepted and published by the IEEE or is withdrawn by the author(s) before acceptance by the IEEE, the foregoing copyright transfer shall be null and void. In this case, IEEE will retain a copy of the manuscript for internal administrative/record-keeping purposes.
5. For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
6. The author hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the author has obtained any necessary permissions. Where necessary, the author has obtained all third party permissions and consents to grant the license above and has provided copies of such permissions and consents to IEEE

You have indicated that you DO wish to have video/audio recordings made of your conference presentation under terms and conditions set forth in "Consent and Release."

CONSENT AND RELEASE

1. In the event the author makes a presentation based upon the Work at a conference hosted or sponsored in whole or in part by the IEEE, the author, in consideration for his/her participation in the conference, hereby grants the IEEE the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive, in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IEEE and live or recorded broadcast of the Presentation during or after the conference.
2. In connection with the permission granted in Section 1, the author hereby grants IEEE the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IEEE from any claim based on right of privacy or publicity.

BY TYPING IN YOUR FULL NAME BELOW AND CLICKING THE SUBMIT BUTTON, YOU CERTIFY THAT SUCH ACTION CONSTITUTES YOUR ELECTRONIC SIGNATURE TO THIS FORM IN ACCORDANCE WITH UNITED STATES LAW, WHICH AUTHORIZES ELECTRONIC SIGNATURE BY AUTHENTICATED REQUEST FROM A USER OVER THE INTERNET AS A VALID SUBSTITUTE FOR A WRITTEN SIGNATURE.

Mohammed Alkhowaiter	25-10-2022
Signature	**Date (dd-mm-yyyy)**

Information for Authors

AUTHOR RESPONSIBILITIES

The IEEE distributes its technical publications throughout the world and wants to ensure that the material submitted to its publications is properly available to the readership of those publications. Authors must ensure that their Work meets the requirements as stated in section 8.2.1 of the IEEE PSPB Operations Manual, including provisions covering originality, authorship, author responsibilities and author misconduct. More information on IEEE's publishing policies may be found at http://www.ieee.org/publications_standards/publications/rights/authorrightsresponsibilities.html Authors are advised especially of IEEE PSPB Operations Manual section 8.2.1.B12: "It is the responsibility of the authors, not the IEEE, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it." Authors are also advised of IEEE PSPB Operations Manual section 8.1.1B: "Statements and opinions given in work published by the IEEE are the expression of the authors."

RETAINED RIGHTS/TERMS AND CONDITIONS

- Authors/employers retain all proprietary rights in any process, procedure, or article of manufacture described in the Work.
- Authors/employers may reproduce or authorize others to reproduce the Work, material extracted verbatim from the Work, or derivative works for the author's personal use or for company use, provided that the source and the IEEE copyright notice are indicated, the copies are not used in any way that implies IEEE endorsement of a product or service of any employer, and the copies themselves are not offered for sale.
- Although authors are permitted to re-use all or portions of the Work in other works, this does not include granting third-party requests for reprinting, republishing, or other types of re-use.The IEEE Intellectual Property Rights office must handle all such third-party requests.
- Authors whose work was performed under a grant from a government funding agency are free to fulfill any deposit mandates from that funding agency.

AUTHOR ONLINE USE

- **Personal Servers**. Authors and/or their employers shall have the right to post the accepted version of IEEE-copyrighted articles on their own personal servers or the servers of their institutions or employers without permission from IEEE, provided that the posted version includes a prominently displayed IEEE copyright notice and, when published, a full citation to the original IEEE publication, including a link to the article abstract in IEEE Xplore. Authors shall not post the final, published versions of their papers.
- **Classroom or Internal Training Use.** An author is expressly permitted to post any portion of the accepted version of his/her own IEEE-copyrighted articles on the author's personal web site or the servers of the author's institution or company in connection with the author's teaching, training, or work responsibilities, provided that the appropriate copyright, credit, and reuse notices appear prominently with the posted material. Examples of permitted uses are lecture materials, course packs, e-reserves, conference presentations, or in-house training courses.
- **Electronic Preprints.** Before submitting an article to an IEEE publication, authors frequently post their manuscripts to their own web site, their employer's site, or to another server that invites constructive comment from colleagues. Upon submission of an article to IEEE, an author is required to transfer copyright in the article to IEEE, and the author must update any previously posted version of the article with a prominently displayed IEEE copyright notice. Upon publication of an article by the IEEE, the author must replace any previously posted electronic versions of the article with either (1) the full citation to the

IEEE work with a Digital Object Identifier (DOI) or link to the article abstract in IEEE Xplore, or (2) the accepted version only (not the IEEE-published version), including the IEEE copyright notice and full citation, with a link to the final, published article in IEEE Xplore.

Questions about the submission of the form or manuscript must be sent to the publication's editor.
Please direct all questions about IEEE copyright policy to:
IEEE Intellectual Property Rights Office, copyrights@ieee.org, +1-732-562-3966

LIST OF REFERENCES

[1] N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[2] M. Alkhowaiter, K. Almubarak, M. Alyami, A. Alghamdi, and C. Zou. Image authentication using self-supervised learning to detect manipulation over social network platforms. In *MILCOM 2022-2022 IEEE Military Communications Conference (MILCOM)*, pages 672–678. IEEE, 2022.

[3] M. Alkhowaiter, K. Almubarak, M. Alyami, A. Alghamdi, and C. Zou. Image authentication using self-supervised learning to detect manipulation over social network platforms. In *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*, pages 672–678, 2022.

[4] M. Alkhowaiter, K. Almubarak, and C. Zou. Evaluating perceptual hashing algorithms in detecting image manipulation over social media platforms. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 149–156. IEEE, 2022.

[5] M. Alkhowaiter, K. Almubarak, and C. Zou. Evaluating perceptual hashing algorithms in detecting image manipulation over social media platforms. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 149–156, 2022.

[6] M. Alkhowaiter, K. Almubarak, and C. Zou. Evaluating perceptual hashing algorithms in detecting image manipulation over social media platforms. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, pages 149–156, 2022.

[7] M. Alkhowaiter, H. Kholidy, M. A. Alyami, A. Alghamdi, and C. Zou. Adversarial-aware deep learning system based on a secondary classical machine learning verification approach. *Sensors*, 23(14), 2023.

[8] M. Alyami, I. Alharbi, C. Zou, Y. Solihin, and K. Ackerman. Wifi-based iot devices profiling attack based on eavesdropping of encrypted wifi traffic. In *2022 IEEE 19th Annual Consumer Communications Networking Conference (CCNC)*, pages 385–392, 2022.

[9] M. Alyami, M. Alkhowaiter, M. A. Ghanim, C. Zou, and Y. Solihin. Mac-layer traffic shaping defense against wifi device fingerprinting attacks. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7, 2022.

[10] G. Apruzzese and M. Colajanni. Evading botnet detectors based on flows and random forest with adversarial samples. In *2018 IEEE 17th International Symposium on Network Computing and Applications (NCA)*, pages 1–8. IEEE, 2018.

[11] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. Manjunath. Exploiting spatial structure for localizing manipulated image regions. In *Proceedings of the IEEE international conference on computer vision*, pages 4970–4979, 2017.

[12] E. Blasch and P. Hanselman. Information fusion for information superiority. In *Proceedings of the IEEE 2000 National Aerospace and Electronics Conference. NAECON 2000. Engineering Tomorrow (Cat. No.00CH37093)*, pages 290–297, 2000.

[13] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[14] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[15] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[16] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[18] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 699–708, 2020.

[19] G. Cohen, G. Sapiro, and R. Giryes. Detecting adversarial samples using influence functions and nearest neighbors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14453–14462, 2020.

[20] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[21] J. Cui and S. Rao. Us army big data military applications and reflections. In *Proceedings of the 2021 3rd International Conference on Big-Data Service and Intelligent Computation*, BDSIC '21, page 92–96, New York, NY, USA, 2022. Association for Computing Machinery.

[22] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[25] L. Du, A. T. Ho, and R. Cong. Perceptual hashing for image authentication: A survey. *Signal Processing: Image Communication*, 81:115713, 2020.

[26] L. Du, A. T. S. Ho, and R. Cong. Perceptual hashing for image authentication: A survey. *Signal Process. Image Commun.*, 81, 2020.

[27] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[28] D. Evon. Hey 'crypto bro's,' that mcdonald's billboard is fake, 2022.

[29] W. Falcon. Pytorch lightning, 2019.

[30] P. Ferrara, T. Bianchi, A. De Rosa, and A. Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012.

[31] A. Gionis, P. Indyk, R. Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, pages 518–529, 1999.

[32] A. Gionis, P. Indyk, R. Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, pages 518–529, 1999.

[33] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[34] A. Gupta, H. Lamba, et al. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 729–736. International World Wide Web Conferences Steering Committee / ACM, 2013.

[35] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736, 2013.

[36] J. Hao, Z. Zhang, S. Yang, D. Xie, and S. Pu. Transforensics: image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15055–15064, 2021.

[37] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[38] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] S. Heller, L. Rossetto, and H. Schuldt. The ps-battles dataset-an image collection for image manipulation detection. *arXiv preprint arXiv:1804.04866*, 2018.

[40] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[41] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[42] K. Hua, Yiwen et al. Holopix50k: A large-scale in-the-wild stereo image dataset. *arXiv preprint arXiv:2003.11172*, 2020.

[43] Y. Hua, P. Kohli, P. Uplavikar, A. Ravi, S. Gunaseelan, J. Orozco, and E. Li. Holopix50k: A large-scale in-the-wild stereo image dataset. *arXiv preprint arXiv:2003.11172*, 2020.

[44] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[45] A. Inc.". "csam detection", 2021.

[46] X. Jia, X. Wei, X. Cao, and H. Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6084–6092, 2019.

[47] H. Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[48] N. I. Korsunov and D. A. Toropchin. Recognition method of near-duplicate images based on the perceptual hash and image key points using. In *2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 261–264. IEEE, 2015.

[49] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[51] W. Kun, L. Tong, and X. Xiaodan. Application of big data technology in scientific research data management of military enterprises. *Procedia Computer Science*, 147:556–561, 2019. 2018 International Conference on Identification, Information and Knowledge in the Internet of Things.

[52] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[53] G. M. Leal, I. Zacarias, J. M. Stocchero, and E. P. d. Freitas. Empowering command and control through a combination of information-centric networking and software defined networking. *IEEE Communications Magazine*, 57(8):48–55, 2019.

[54] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[55] Y. Li, D. Wang, and L. Tang. Robust and secure image fingerprinting learned by neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):362–375, 2019.

[56] A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[57] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[58] K. Ma, H. Zhang, R. Wang, and Z. Zhang. Target tracking system for multi-sensor data fusion. In *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pages 1768–1772, 2017.

[59] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[60] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[61] S. McSweeney". "our ongoing approach to the war in ukraine", 2022.

[62] Meta. Facebook's third-party fact-checking program, 2022.

[63] D. Mikkelson. Does this photograph show president bush reading a book upside-down?, 2002.

[64] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[65] H. Müller, P. Clough, T. Deselaers, B. Caputo, and I. CLEF. Experimental evaluation in visual information retrieval. *The Information Retrieval Series*, 32:1–554, 2010.

[66] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020.

[67] A. Novozamsky, B. Mahdian, and S. Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 71–80, 2020.

[68] K. Olafson and T. Tran". "social media image sizes 2021: Cheat sheet for every network", 2022.

[69] K. M. I. Omid Jafari et al. A survey on locality sensitive hashing algorithms and their applications. *CoRR*, abs/2102.08942, 2021.

[70] N. Papernot and P. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

[71] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[72] A. Paszke, S. Gross, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[74] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

[75] C. Qin, X. Chen, J. Dong, and X. Zhang. Perceptual image hashing with selective sampling for salient structure features. *Displays*, 45:26–37, 2016.

[76] C. Qin, X. Chen, D. Ye, J. Wang, and X. Sun. A novel image hashing scheme with perceptual robustness using block truncation coding. *Information Sciences*, 361:84–99, 2016.

[77] C. Qin, E. Liu, G. Feng, and X. Zhang. Perceptual image hashing for content authentication based on convolutional neural network with multiple constraints. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4523–4537, 2021.

[78] M. H. J. R. Venkatesan, S.-M. Koon and P. Moulin. Robust image hashing. *Proceedings 2000 International Conference on Image Processing*, 3:664–666, 2000.

[79] T. Reddy, S. P. RM, M. Parimala, C. L. Chowdhary, S. Hakak, W. Z. Khan, et al. A deep neural networks based model for uninterrupted marine environment monitoring. *Computer Communications*, 157:64–75, 2020.

[80] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[81] Rettore, P. H. L., G. Maia, L. A. Villas, and A. A. F. Loureiro. Vehicular data space: The data point of view. *IEEE Communications Surveys Tutorials*, 21(3):2392–2418, thirdquarter 2019.

[82] Rettore, Paulo H. L. , B. Pereira, R. Rigolin F. Lopes, G. Maia, L. Villas, and A. Loureiro. Road data enrichment framework based on heterogeneous data fusion for its. *IEEE Transactions on Intelligent Transportation Systems*, 01 2020.

[83] Y. Roth and N. Pickles". "updating our approach to misleading information", 2020.

[84] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[85] V. B. S. and R. V. Babu. Single-step adversarial training with dropout scheduling, 2020.

[86] F. Sabahi, M. O. Ahmad, and M. Swamy. Content-based image retrieval using perceptual image hashing and hopfield neural network. In *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, pages 352–355. IEEE, 2018.

[87] P. Samangouei, M. Kabkab, and R. Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

[88] B. P. Santos, Rettore, P. H. L., H. S. Ramos, L. F. M. Vieira, and A. A. F. Loureiro. Enriching traffic information with a spatiotemporal model based on social media. In *2018 IEEE Symposium on Computers and Communications (ISCC)*, pages 00464–00469, June 2018.

[89] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[90] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[91] B. Steele II, H. A. Kholidy, et al. 5g networks security: Attack detection using the j48 and the random forest tree classifiers. *SUNY Polytechnic Institute*, 2020.

[92] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, et al. Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. *arXiv preprint arXiv:2211.06318*, 2022.

[93] C. Strobl, J. Malley, and G. Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323, 2009.

[94] L. Struppek, D. Hintersdorf, D. Neider, and K. Kersting. Learning to break deep perceptual hashing: The use case neuralhash. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 58–69, 2022.

[95] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[96] M. N. "T. Chen, S. Kornblith and G. Hinton". "a simple framework for contrastive learning of visual representations". *"In International conference on machine learning"*, "10":"1597–1607", "2020".

[97] Y. Tang, Zhenjun et al. Robust image hashing with dominant dct coefficients. *Optik - International Journal for Light and Electron Optics*, 125(18):5102–5107, 2014.

[98] Z. Tang, H. Lao, X. Zhang, and K. Liu. Robust image hashing via dct and lle. *Computers & Security*, 62:133–148, 2016.

[99] Z. Tang, X. Zhang, X. Li, and S. Zhang. Robust image hashing with ring partition and invariant vector distance. *IEEE transactions on information forensics and security*, 11(1):200–214, 2015.

[100] Z. Tang, X. Zhang, X. Li, and S. Zhang. Robust image hashing with ring partition and invariant vector distance. *IEEE Transactions on Information Forensics and Security*, 11(1):200–214, 2016.

[101] U. University". "usc-sipi image database", 1977.

[102] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[103] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[104] K. Wang, Xiaofeng et al. A visual model-based perceptual image hash for content authentication. *IEEE Transactions on Information Forensics and Security*, 10(7):1336–1349, 2015.

[105] X. Wang, K. Pang, X. Zhou, Y. Zhou, L. Li, and J. Xue. A visual model-based perceptual image hash for content authentication. *IEEE Transactions on Information Forensics and Security*, 10(7):1336–1349, 2015.

[106] Y. Wang, F. Tahmasbi, J. Blackburn, B. Bradlyn, E. De Cristofaro, D. Magerman, S. Zannettou, and G. Stringhini. Understanding the use of fauxtography on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 776–786, 2021.

[107] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020.

[108] F. Waseda, S. Nishikawa, T.-N. Le, H. H. Nguyen, and I. Echizen. Closer look at the transferability of adversarial examples: How they fool different models differently. In *Proceedings*

of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1360–1368, 2023.

[109] T. website". "what is tineye", 2022.

[110] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler. Coverage—a novel database for copy-move forgery detection. In *2016 IEEE international conference on image processing (ICIP)*, pages 161–165. IEEE, 2016.

[111] Y. Wu, W. AbdAlmageed, and P. Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9543–9552, 2019.

[112] L.-l. Xu and F. Jin. Research on military intelligence value evaluation method based on big data analysis. In Y.-D. Zhang, S.-H. Wang, and S. Liu, editors, *Multimedia Technology and Enhanced Learning*, pages 192–200, Cham, 2020. Springer International Publishing.

[113] C.-P. Yan, C.-M. Pun, and X.-C. Yuan. Quaternion-based image hashing for adaptive tampering localization. *IEEE Transactions on Information Forensics and Security*, 11(12):2664–2677, 2016.

[114] C.-P. Yan, C.-M. Pun, and X.-C. Yuan. Quaternion-based image hashing for adaptive tampering localization. *IEEE Transactions on Information Forensics and Security*, 11(12):2664–2677, 2016.

[115] C. "Zauner. "phash.org: Home of phash, the open source perceptual hash library", 2010.

[116] Y. Zhao, S. Wang, X. Zhang, and H. Yao. Robust hashing for image authentication using zernike moments and local features. *IEEE transactions on information forensics and security*, 8(1):55–63, 2012.

[117] W. Zhen-kun, Z. Wei-Zong, L. Peng-fei, D. Yi-hua, Z. Meng, G. Jin-hua, et al. A robust and discriminative image perceptual hash algorithm. In *2010 Fourth International Conference on Genetic and Evolutionary Computing*, pages 709–712. IEEE, 2010.

[118] Y. Zheng". "casia dataset", 2021.

[119] Y. Zheng". "casia dataset", 2021.

[120] Z. Zhu, Wen et al. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland*, 19:67, 2010.

[121] P. Zißner, P. H. L. Rettore, B. P. Santos, J. F. Loevenich, and R. R. F. Lopes. Datafits: A heterogeneous data fusion framework for traffic and incident prediction. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–0, 2023.