NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PREDICTING THE UNKNOWN: MACHINE LEARNING TECHNIQUES FOR VIDEO FINGERPRINTING ATTACKS OVER TOR

by

Elissa S. Kim

December 2021

Thesis Advisor: Armon C. Barton
Second Reader: Mathias N. Kolsch

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2021	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE PREDICTING THE UNKNOWN: MACHINE LEARNING TECHNIQUES FOR VIDEO FINGERPRINTING ATTACKS OVER TOR	**5. FUNDING NUMBERS**
6. AUTHOR(S) Elissa S. Kim	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

In recent years, anonymization services such as Tor have become a popular resource for terrorist organizations and violent extremist groups. These adversaries use Tor to access the Dark Web to distribute video media as a way to recruit, train, and incite violence and acts of terrorism worldwide. This research strives to address this issue by examining and analyzing the use and development of video fingerprinting attacks using deep learning models. These high-performing deep learning models are called Deep Fingerprinting, which is used to predict video patterns with high accuracy in a closed-world setting. We pose ourselves as the adversary by passively observing raw network traffic as a user downloads a short video from YouTube. Based on traffic patterns, we can deduce what video the user was streaming with higher accuracy than previously obtained. In addition, our results include identifying the genre of the video. Our results suggest that an adversary may predict the video a user downloads over Tor with up to 83% accuracy, even when the user applies additional defenses to protect online privacy. By comparing different Deep Fingerprinting models with one another, we can better understand which models perform better from both the attacker and user's perspective.

14. SUBJECT TERMS Tor, deep learning, convolutional neural networks, Dark Web, classifier, features, closed-world, defense models, accuracy	**15. NUMBER OF PAGES** 59
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

PREDICTING THE UNKNOWN: MACHINE LEARNING TECHNIQUES FOR VIDEO FINGERPRINTING ATTACKS OVER TOR

Elissa S. Kim
Lieutenant, United States Navy
BS, United States Naval Academy, 2014

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN CYBER SYSTEMS AND OPERATIONS

from the

NAVAL POSTGRADUATE SCHOOL
December 2021

Approved by: Armon C. Barton
Advisor

Mathias N. Kolsch
Second Reader

Alex Bordetsky
Chair, Department of Information Sciences

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In recent years, anonymization services such as Tor have become a popular resource for terrorist organizations and violent extremist groups. These adversaries use Tor to access the Dark Web to distribute video media as a way to recruit, train, and incite violence and acts of terrorism worldwide. This research strives to address this issue by examining and analyzing the use and development of video fingerprinting attacks using deep learning models. These high-performing deep learning models are called Deep Fingerprinting, which is used to predict video patterns with high accuracy in a closed-world setting. We pose ourselves as the adversary by passively observing raw network traffic as a user downloads a short video from YouTube. Based on traffic patterns, we can deduce what video the user was streaming with higher accuracy than previously obtained. In addition, our results include identifying the genre of the video. Our results suggest that an adversary may predict the video a user downloads over Tor with up to 83% accuracy, even when the user applies additional defenses to protect online privacy. By comparing different Deep Fingerprinting models with one another, we can better understand which models perform better from both the attacker and user's perspective.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

API	Application Programming Interface
CAPTCHA	Completely Automated Public Turing Test
CNN	Convolutional Neural Network
DF	Deep Fingerprinting
DL	Deep Learning
DOD	Department of Defense
ID	Identification
ISP	Internet Service Provider
ML	Machine Learning
NPS	Naval Postgraduate School
PCAP	Packet Capture
TCP	Transmission Control Protocol
Tor	The Onion Router
VF	Video Fingerprinting
VM	Virtual Machine
Wi-Fi	Wireless Fidelity
WF	Website Fingerprinting
W-T	Walkie-Talkie
WTF-PAD	Website Traffic Fingerprinting Protection with Adaptive Defense

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

First, I would like to thank my thesis advisor, Dr. Armon Barton, for his continued guidance and insight throughout my research as well as giving me the opportunity to learn more about the fascinating realm of machine learning. I would also like to thank Dr. Mathias Kölsch as my second reader for his thoughtful input and recommendations. Finally, thank you to all my friends and family that have believed in me from the start and provided me the love, support, and motivation to keep going.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. INTRODUCTION

More than ever before, cybersecurity has escalated to the forefront of priorities for national security, and for the protection of the country's entire critical infrastructure. The Department of Defense (DOD) and other agencies have recognized the urgent need for robust cyber related knowledge, skills, resources, and firepower to defend against bad actors who exist across a whole spectrum of threat levels. Even worse, it is often extremely difficult to discern the origin of these cyberattacks and the actual capabilities of these malicious cyber actors.

Many of these adversaries hide behind the walls of encrypted anonymous networks such as The Onion Router (Tor) Network. Tor is an anonymous communication system that has recently gained substantial popularity serving millions of users world-wide. In [1], Lewman states that Tor's user base spans a wide range of demographics; a number of these users consist of journalists, political dissidents or "whistleblowers," law enforcement agencies, and other regular users who simply desire for more privacy online. More importantly, today's modern adversaries and terrorists are also utilizing the Tor network to access the Dark Web in order to facilitate their activities and evade monitoring and tracking online. According to Sabbah et al. [2], Tor's unique internet architecture provides a quick, easy-access domain for extremist groups to preserve their anonymity. The Dark Web continues to be a valuable tool for numerous terrorist platforms to communicate, fundraise, spread propaganda, steal information and data, and more—all in complete secrecy. Recently, recruitment, instruction, and training videos have become a more prominent threat on the international stage by encouraging the proliferation and exposure of violent extremism. Salem et al. [3]. determined that the dissemination of these digital materials allows extremist groups to support their goals of spreading their ideologies and influencing potential recruits all around the world.

An emerging area to improve automated capability to disrupt such terrorist recruitment efforts and related activities is via website fingerprinting (WF) attacks [4]. This

type of cyber-attack helps determine the subject of a website or other digital material a user is viewing based on collected network traffic. To perform a WF attack, an attacker will passively eavesdrop on a client's web activity who is using low-latency anonymity networks such as proxies and virtual protected networks (VPN) [5]. The attacker will then leverage certain features of the network packet sequence such as time, direction, and size. Now, the attacker has the *fingerprints* of the webpage and can use various machine learning (ML) techniques to classify the packet sequence to determine the particular website the client visited.

As Sirinam et al. [4]. point out, "state-of-the-art website fingerprinting attacks have been shown to be effective even against Tor." Consequently, there is growing importance for the DOD and other federal agencies to shift towards developing and integrating new ML techniques and cyber defense strategies to combat terrorists who frequently utilize the Dark Web for threatening, harmful activities especially for information sharing and mobilizing resources via online avenues. By gaining a better understanding of these unique fingerprinting attack methods, we are better equipped to disrupt terrorist and malicious cyber activity in addition to devising new defense mechanisms to protect ourselves from these attacks.

Similar to WF attacks, video fingerprinting (VF) attacks over Tor involves passively observing a victim's network traffic packets without any modification to the client and server. However, rather than determining what web page was visited, VF allows the attacker to infer what video is being streamed by the client. As pointed out by Lu [6], VF has become "an essential and enabling tool adopted by the industry for video content identification and management in online video distribution." However, in the current literature, there is an absence of research toward studying the effects of applying VF over the Tor Network. This research aims to utilize VF methods to address the growing popularity of video streaming amongst terrorist organizations and violent extremist groups over anonymous networks such as Tor.

Following previous work from Sirinam et al. [4], Schuster et al. [7], and Campuzano [8], we will explore applying a deep learning (DL) classifier to predict the video a user watches over Tor as well as the genre of the video. Video genre refers to the

general category a video may be characterized as such as a movie trailer or music video. Our research utilizes a new type of attack called Deep Fingerprinting (DF) proposed by Sirinam et al. [4]. that will be used with DL models to produce high accuracy scores for predicting video streams based on raw network traffic. By impersonating an attacker, we eavesdrop on a client's traffic transmission and collect datasets composed of these network traces to train the DL classifier. We attempt to discover the effectiveness of various attack models using these DL classifiers to determine which model achieves the highest video prediction accuracy score.

B. PURPOSE AND SCOPE

This research serves to explore a simulated video fingerprinting (VF) attack model in which we behave as the attacker who is passively observing Tor traffic in between the user and their first hop within the Tor network. We strive to achieve the highest video prediction accuracy by unmasking a user's video download and accurately identifying the short video streams over an encrypted network to better understand our adversaries and their capabilities. Large quantities of video traffic traces that include packet time, direction, and size will be collected to train our deep learning classifier DF that used a similar concept that Sirinam et al. [4]. applied to their WF attack model. This research will further explore (DL) techniques proposed by Sirinam et al. and Campuzano [4], [8], who both use convolutional neural network (CNN) models and VF data gathered from the video streaming website, YouTube, to train the models to attain high classification accuracy using raw traffic data. This process will allow these DL models to accurately detect unusual frequency patterns of videos as well as correlate video traffic patterns in various attack settings.

Lastly, several different DL models will be compared to assess which model performs the best for accurately predicting a video that was watched over Tor. By collecting a high volume of video traffic packet data, we are able to apply DF methods to our research to determine the effectiveness of VF attacks in different attack settings and hyperparameters.

C. THESIS ORGANIZATION

The remaining chapters are organized as follows. Chapter II provides background information on the Dark net or "Dark Web" and an overview of the Tor network. The chapter then explains the VF process and DL models used to achieve results and findings to include previous related work done in the same area of research. Chapter III describes the methodology used throughout the study to obtain data using a video crawler program as well as further computation and analysis of the datasets. Chapter IV describes the implemented DL models and their hyperparameters in further detail, the simulated closed-world setting used to evaluate the effectiveness of each model, and their results. Lastly, Chapter V summarizes the contributions towards this area of research, possible recommendations for future work, and our conclusion.

II. BACKGROUND AND RELATED WORK

This chapter provides both background and comprehensive review of the literature related to my research. The first and second sections provide general and detailed information of the Dark net or "Dark Web" and the Tor network as well as its application and framework. The third section discusses WF and its relevance and value as a common cyber tool used by adversaries to identify video streams viewed over an encrypted network such as Tor. The fourth section explains VF and its application throughout this research. The fifth section highlights the DL methods and how they may be employed to make similar human-like predictions and decisions to be used in a simulated VF attack. The sixth and seventh sections serve to provide more detail of the machine and DL models used throughout this research. The last section provides insight to previous work done in this area of study.

A. DARK NET

Dark net, or commonly known as the "Dark Web," may be conceptualized as a collection of sites that utilize "special routing systems…designed to provide anonymity for both visitors to websites and publishers of these sites" [9]. In [9], Gehl favors a technical definition of the Dark Web describing it as "websites built with standard web technologies (HTML, CSS, server-side scripting languages, hosting software) that can be viewed with a standard web browser…which is routed through special routing software packages." Thus, the important distinction between the regular internet and the Dark Web is the ability to achieve anonymity through the Dark Web medium. The Dark Web architecture uses anonymity or *hidden* service tools to conceal IP addresses and encrypt traffic data [10]. It is part of the internet that "is not a separate physical network but an application and protocol layer riding on existing networks" [10].

The Dark Web is notorious for being the hotbed of criminal, illicit activity while serving as the neural pathway into the deep recesses of the internet. As Samtani et al. [11] mentions in their research, malicious cyber actors are able to leverage the privacy afforded by the Dark Web to procure various attack tools such as malware, remote administration

tools (RATs), ransomware, botnets, spyware, and other emerging threats. However, the Dark Web is also frequented by journalists, political dissidents or "whistleblowers," law enforcement agencies, and other regular users who also desire to achieve anonymity online. This obscure, complex domain is characterized as being a highly concealed portion of the internet that few users will ever interact with or view. In fact, "when performing a regular search on the internet, what is returned really only makes up less than one percent of all the information that is actually out there" [12].

Websites that are accessible through this medium are not conventionally indexed and can only be accessed via specialized web browsers. As a decentralized network of internet websites, the Dark Web allows users to become incognito by routing their online traffic through multiple servers and encrypting these communications from beginning to end [3]. Thus, the Dark Web often serves as an ideal hideaway for adversaries to evade attribution and detection. Although many major terrorist organizations utilize social media applications such as Facebook and Twitter [13] as a more public display of their strategies, they turn to the Dark Web to use its encrypted channels to communicate and secretly orchestrate their attacks. Amongst numerous extremist and terrorist organizations, Dark Web "online forums have also facilitated the 'leaderless resistance' movement, a decentralized and diffused tactic that has made it increasingly difficult for law enforcement officials to detect potentially violent extremists" [13].

By studying the Dark Web in greater detail, we are able to better understand its relevance to cyber security concerns as well as preventing attacks and malicious activity and services [14]. The rapid globalization and ubiquitous access to the internet continues to allow illegal online activities and acts of terrorism to become more prevalent especially through the Dark Web medium. Today, it has become apparent that terrorist and extremist groups continue to grow their knowledge base and proficiency with the Dark Web in order to find new ways to utilize anonymous networks while remaining undetected [14]. As we remember past catastrophic events such as 9/11, it becomes increasingly more essential to intercept terrorist communications and networks throughout these hidden channels to prevent future attacks and safeguard national security.

B. THE ONION ROUTER "TOR"

Tor is an open-source software that enables access to the Dark Web and supports anonymous communications. It was originally developed by the United States Navy in the mid-1990s to enhance privacy and protect online communications and identities within the U.S. intelligence community [15]. Today, the Tor network is the most widely used anonymity system and comprises approximately "7,000 relays or proxies which together carry terabytes of traffic every day" [16]. Tor can essentially be understood as a browser that utilizes a virtual protected network (VPN) proxy connection with a default search engine similar to Google or Bing which can be used to access the Dark Web via randomly chosen proxies or relay points spread throughout the world. This allows for an encrypted circuit path to be established between the client and the destination. Tor can also be used to access any normal website; the majority of Tor's bandwidth is devoted to users accessing normal websites through the Tor network. Our adversary model makes this assumption as the adversary is interested in users accessing YouTube videos over Tor.

During the first step of the Tor process, the user makes a request to the Tor directory which consists of numerous relay nodes. By default, three random nodes are selected (guard "entry," middle, and exit) in order to establish symmetric session keys and a circuit [17]. The user's data is bundled into encrypted packets while being routed through the various relay nodes or onion routers (OR) on its way to the ultimate destination. The encrypted packets, or *cells*, form Tor's multi-layered routing infrastructure in which each OR "maintains a TLS connection to every other onion router" [17]. As Dingledine et al. [17] explain, each cell consists of a header and payload in which the header "includes a circuit identifier (circID) that specifies which circuit the cell refers to (many circuits can be multiplexed over the single TLS connection) and a command to describe what to do with the cell's payload."

Cells are categorized as either *control* cells or *relay* cells serving different purposes [17]. Control cells consist of several different commands such as "*padding* (currently used for keepalive, but also usable for link padding); *create* or *created* (used to set up a new circuit); and *destroy* (to tear down a circuit)" [17]. Relay cells are used to "carry end-to-end stream data" [17]. As Haraty and Zantout [18] describe, "each cell is encrypted/

decrypted at every node and each node can only reveal a single encrypted layer in a cell" using the symmetric session key. Once the data that originated from the user is sent through the last node via cell encapsulation and the encryption key methodology, the ultimate destination is presented with the revealed data without the identity of the user [18]. The data will then transit backwards along the same path in which the last node "has to prepare the same encapsulated set of layers in an onion similar to the one the client has prepared earlier using the reverse order of layers originally sent by the client" [18]. Figure 1 is a graphical representation of the Tor network structure and its basic components.

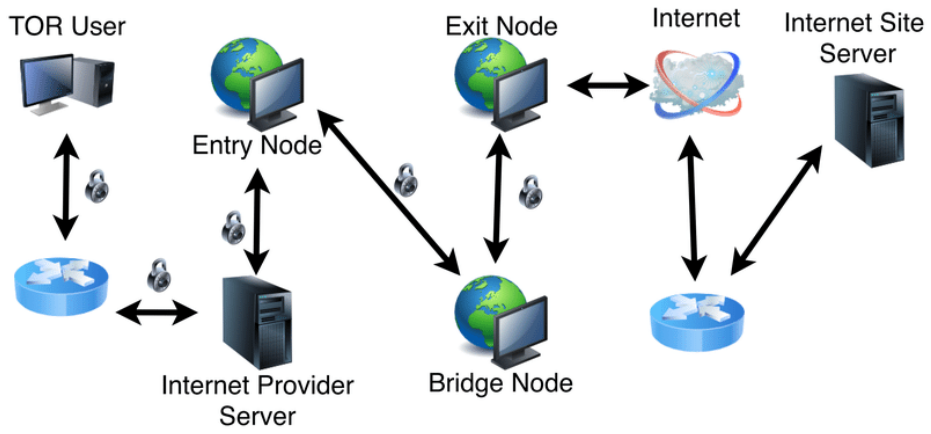

Figure 1. Tor network operation. Adapted from [19].

Despite Tor's intricate layered or "onion" routing infrastructure, "the designers of Tor admit that the anonymous network does not prevent against the global adversaries that have exclusive network/resource access and are capable of monitoring traffic on all networks their users are connected to" [18]. Although the data packet in the final path does not include the identity of the client, numerous studies have been conducted to discover how an adversary can still use other information to uncover visited websites over a Tor connection. Gilad and Herzberg [20] mentioned that the low-latency nature of Tor makes it vulnerable to traffic correlation attacks by an eavesdropper who may passively observe traffic between client and guard, and simultaneously between her exit and destination server. Dingledine [17] adds, "because these designs typically involve many packets that

must be delivered quickly, it is difficult for them to prevent an attacker who can eavesdrop both ends of the communication from correlating the timing and volume of traffic entering the anonymity network with traffic leaving it".

C. WEBSITE FINGERPRINTING

In recent years, "research shows that website fingerprinting (WF) is a growing threat to privacy-sensitive web users, especially when using machine learning techniques such as deep learning or machine learning (DL/ML) to attack website fingerprint, reducing the effectiveness of the previous defense strategies" [21]. WF attacks are passive traffic analysis attacks designed to recognize traffic patterns of specific websites by observing the TCP header information within the network traffic. Although encrypting tunnels are capable of hiding content and addresses between client and web server, packet information such as size, time, and direction can still be seen by outsiders [16]. The adversary can successfully carry out this attack by simply eavesdropping and extracting network traces from unsuspecting Tor users, then matching those traces to a collection of pre-recorded website fingerprints to identify which website the user visited [22].

Several different approaches have been developed to advance WF attack techniques using ML and DL classifiers and algorithms. Certain features, or individual independent variables or characteristics are used to make predictions through pattern recognition and ML processes. Hermann et al. [23]. presented a novel WF technique based on a Multinomial Naïve-Bayes classifier in which text mining was used to normalize frequency distribution of observable IP packet sizes. By observing and conducting statistical analysis of packet sizes sent from client to server, they were able to correctly identify "up to 97% of requests on a sample of 775 sites and over 300,000 real-world traffic dumps recorded over a two-month period" [23].

Other approaches utilized different classifiers, variables, and settings. Panchenko et al. [24]. specifically focused on traffic time, direction, and volume for features to be used in their WF attack models as well as using support vector machines. In addition, they introduced both closed and open-world settings where closed-world assumes that the attacker knows all the web pages in advance whereas the open-world scenario is the

opposite: the attacker is not privy to any information regarding which web pages are visited by the user ahead of time. This new approach resulted in improved recognition that increased from 3% to 55% using closed-world settings and achieved "a surprisingly high true positive rate of up to 73% for a false positive of 0.05%" [24] for the more complex open-world scenario.

D. VIDEO FINGERPRINTING

Due to the fast-growing popularity of video streaming on the internet, VF methods have become essential for video content identification, management, and distribution. According to Rao et al. [25], "recent studies have shown that video streaming is responsible for 25%-40% of all internet traffic." Similar to WF, VF may also be used over an encrypted network where the same adversary can build his own database of video traffic to correlate with what he observes between client and Tor entry node. Despite video content being hidden using transport-layer encryption, network characteristics such as burst patterns can still be discerned and can be tell-tale indicators used to identify which video was streamed [7]. ML and DL classifiers can once again be trained and used to identify encrypted traffic patterns. If trained and tested properly, the adversary will be able to predict what sort of videos clients are watching over Tor with relatively high accuracy. For the purpose of this research, we focused on VF attacks on previously-known videos in order to address the main concern of terrorist/extremist's frequent use of video content and media sharing within the Dark Web.

Today, an internet user has a plethora of video streaming websites to choose from as video streaming has grown exponentially. In [25], Rao et al. mention that "the two dominant sources for video streaming… are Netflix and YouTube." Sandevine [26] reports that video streaming is responsible for "58% of total downstream volume of traffic on the internet…Netflix is 15% of the total downstream volume of traffic across the entire internet" and YouTube reigns as the most highly used video streaming application on mobile devices and shares 11% of the global traffic share.

As Sirinam et al.[4] implemented in their research, we also focused our research on collecting network traces from YouTube over the Tor network. By first extracting and

observing packet characteristics such as transmission time, raw packet sequence, and bursts of traffic, we were able to train our DL classifiers on these features to recognize the videos that the users viewed over their Tor connection. In contrast to WF attacks, where an adversary can only deduce the homepage or hosting site that a user visited, VF attacks have the potential to be significantly more invasive by allowing the adversary to recognize specific video content that is being watched, thus making VF a greater potential threat to user privacy [7].

Following previous research from Sirinam and Juarez et al. [4],[22], our research focuses on closed-world results. A closed-world scenario assumes that the user can only view and download videos that the adversary will use to train his model. A total of three genre types are used in this research with each comprising 9 videos. The adversary is allowed to train on every video in each of the three genres. The user, or victim, is unable to download any other videos using Tor than these selected sets of videos. Recent research shows that DL-based classifiers that operate on raw packet information achieve the best results, with less than 2% error in a closed-world setting [4], [22].

E. MACHINE LEARNING

Machine learning (ML) technology is found in virtually every industry with its numerous different applications. Modern society is constantly driven by an immense amount of data that demands fast computing and analysis to improve efficiency and decision-making. Continuous technological developments in ML have resulted in improved automated capabilities for a wide range of applications and fields of study.

More importantly, recent advancements in ML have proven to be highly effective and useful with encrypted traffic characterizations and analysis [27]. Our research strives to effectively use these ML capabilities to conduct VF traffic analysis attacks to better predict the user's activity. Following previous work from Sirinam et al. and Campuzano [4], [8], our research uses the same DF classifier algorithms and features extracted from raw traffic packet data such as packet direction, size and time to predict what video is being watched and the genre type of the video. The adversary trains a classifier on these features so that it can reliably identify the video being viewed over the Tor network.

F. DEEP LEARNING MODELS

As an important subfield of ML [28], DL is comprised a multi-layer neural network architecture that is programmed to perform different kinds of transformations or predictions based on their inputs. Each layer considers an increasing level of complexity and abstraction where one layer's output is fed as input to the next layer. As Rimmer [28] mentions, "Processing sufficient amounts of representative data enables the deep neural network to not only precisely reveal the identifying features but also generalize better to unseen test instances." According to Sirinam et al. [4]., "while state-of-the art attacks use classifiers that are popular in many applications, *deep learning* (DL) has shown to outperform traditional machine learning techniques in many domains, such as speech recognition, visual object recognition, and object detection."

Moreover, Convolutional Neural Networks (CNNs) have been shown in recent literature and research to outperform other ML models for traffic analysis attacks [4]. Schuster et al. [7]. demonstrated that CNNs are especially useful for identifying patterns in video streams due to their distinguishable traffic burst patterns. Furthermore, DL models utilizing CNNs are able to achieve high accuracy scores on raw traffic with its high performing feature extractors and minimal preprocessing [28]. Unlike other ML models, CNNs do not require manually selecting and fine-tuning of features. Each layer in a CNN is comprised kernels or filters that serve as building blocks of the DL model. These layers are fully connected with one another and use multi-dimensional arrays with input data to produce feature maps as represented in Figure 2 [28].

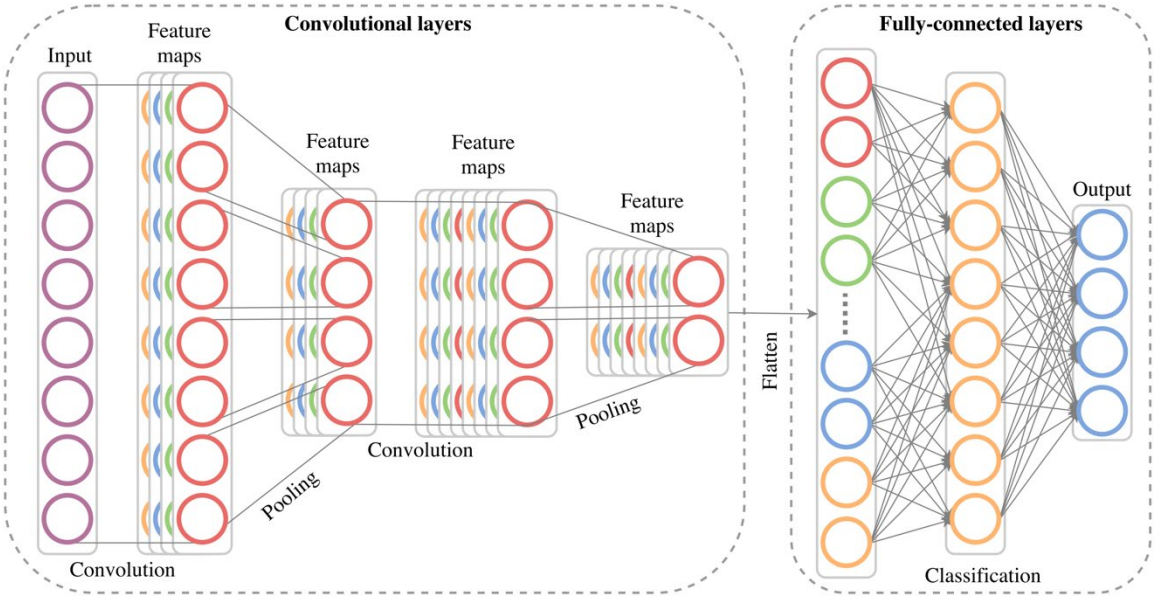

Figure 2. Convolutional Neural Network. Adapted from [28].

As Rimmer et al. explained in [28], "the kernel is applied spatially to small regions of the input, thus enabling sparse connectivity and reducing the actual parameter learning in comparison to fully-connected layers." In our instance, the kernel strives to train on individual parts of a given feature set to yield a usable *fingerprint* from a video traffic trace. During convolutional operations, once each region of input is combined with a filter, intermediate values are produced which then becomes the input to an activation function [4]. As Sirinam et al. [4]. points out, "the output of the activation function is then fed into a pooling layer. The pooling layer progressively reduces the spatial size of the representation from the feature map to reduce the number of parameters and amount of computation." In other words, pooling involves subsampling operations which is often used to identify distinguishable parts of a *fingerprint* in a given trace while disregarding surrounding traffic [28]. Lastly, the final layers will then output the predictions. In Rimmer et al.'s study, they show that "the network can include a whole series of convolution and pooling layers in order to extract more abstract features" [28].

G. DEFENSE MODELS

This research will also introduce lightweight fingerprinting defenses for Tor to degrade our simulated VF attacks using the CNN model: Website Traffic Fingerprinting Protection with Adaptive Defense (WTF-PAD) and Walkie-Talkie (W-T) defenses. We replicate the models that Sirinam and Juarez et al. [4], [22] used for WF attacks. These DF models comprise "more convolutional layers, better protections against overfitting, hyperparameters that vary with the depth of each layer, activation functions tailored to our input format, and a two-layer fully connected classification network" [4]. Both WTF-PAD and W-T were constructed to be low-overhead, flexible, and effective defenses against WF fingerprinting attacks. Our work compares results from these two defense models with the defenseless Tor traffic model called Non-Defended (NoDef).

Following [4], the W-T defense ensures that, "the Tor browser communicate with the web server in *half-duplex* mode, in which the client sends a request (such as for an image file) only after the server has fulfilled all previous requests." Consequently, communications from both the client and server sides become non-overlapping bursts in each direction [4]. As Wang and Goldberg [29] proposed in their research, W-T "is highly effective against all known attacks at overhead costs much lower than all known effective defenses." Dummy packets are also used in W-T; they are fake packets that do not contain real information yet are indistinguishable from real packets to the attacker [4]. They act as cover traffic that "makes WF features less distinctive, thus increasing the rate of classification errors committed by the adversary" [4].

Similar to the W-T defense, WTF-PAD is another countermeasure that alleviates low latency overhead. Juarez et al. [22]. proposed WTF-PAD to be "a system designed for deploying adaptive padding for WF defense in Tor." Adaptive padding serves as a useful defense against timing analysis, and "saves bandwidth by adding the padding only upon low usage of the channel, thus masking traffic bursts and their corresponding features" [4]. Shmatikov and Wang [30] also explain that "the purpose of adaptive padding is to prevent the attacker from determining which of the multiple simultaneous connections is being carried on a given network link" as well as providing "significant protection against active attacks at a relatively low extra latency cost."

In [4], Sirinam et al. achieved 98.3% accuracy for their DF WF attack in a closed-world setting with no defenses. While using the lightweight website defense models, they achieved lower accuracy scores for each showing their effectiveness against WF attacks. WTF-PAD achieved 90.7% accuracy whereas W-T defense was as low as 49.7% [4].

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

This chapter discusses the data collection process of video traffic packet captures over the Tor network using a video crawler within a virtual machine (VM). There is also mention of some errors that were encountered while acquiring traffic packet captures which influenced the number of usable data points for training and testing the selected DL models. In addition, this chapter discusses the various DL models used in greater detail as well as the hyperparameters utilized for each model to better measure video prediction scores.

A. DATA COLLECTION PROCESS

The videos that were selected for our research aimed to possess similar characteristics as the recruitment and propaganda videos that terrorists and other extremist groups may potentially disseminate over the internet. YouTube was used as the source for video downloading under the assumption that these adversaries would likely use YouTube to post their videos to be viewed by regular users.

Video data collection for this research consisted of downloading a variety of videos over Tor in separate batches to include 10 downloads per video from YouTube per batch. Videos are categorized by 3 different genres: The first genre consists of Disney movie trailers, second genre is music videos, and third genre is sports clips. Each genre includes 9 videos of varying lengths (2-4 minutes). The videos are identified by their *video ID* number which we label in sequential order. Disney movie trailers comprise video IDs 0–8, music videos comprise video IDs 9–17, and sports clip videos comprise video IDs 18–26. Each video dataset batch consisted of 90 downloads total: 10 downloads for each of the 9 videos. Thus, the total number of downloads is the same across all three video genres.

The selection of video genre types for our work was based on assumed video characteristics of propaganda videos used by terrorist organizations. Such videos often include loud audio and fast-motion sequences – similar qualities seen in movie trailers, music videos, and sport highlight videos. In contrast, selecting video genres with dissimilar characteristics such as ambient meditation videos would not suit the type of videos we strive to predict on the Dark Web. Following Wang and Goldberg [31], they demonstrated

how added high-bandwidth or *loud* audio streaming noise can affect WF. In [31], they have shown that "the attacker can still classify packet sequences with high levels of noise by finding the noise packets from the client's packet sequence and adding them to his data set…." Based on Wang and Goldberg's [31] findings, we deduce that the high bandwidth rate of our selected videos would still allow us to identify these videos somewhat accurately while closely resembling our real-world videos of interest.

1. Video Crawler

Similar to the Tor browser crawler that Sirinam and Juarez applied in their WF attack models [4,20], we used a modified Tor browser crawler that read YouTube video IDs instead of website URLs. The video crawler developed by Mathews [32] was installed within a VM using Lubuntu version 20.04 operating system with 350 Megabytes of RAM [8]. A Home Wi-Fi connection was used to facilitate each VF attack scenario.

Docker software is used to execute the video crawler inside an encapsulated package called *tbcrawl* within the container. To begin passively downloading the selected videos (this is synonymous to a user clicking *play* for a video) from the Tor network, the crawler uses Selenium to automate the Tor Browser Bundle (TBB) version 8.0.2 and the YouTube Data Application Programming Interface (API) to stop the video capture instance once complete [32]. In addition, we maintain YouTube's automatic adjustment and default stream resolution settings to best simulate average user behavior. The crawler then uses *tcpdump* to collect and analyze the transmitted network packets.

The default configuration file for Tor is set to automatically close its circuits after 10 minutes. Following Wang and Goldberg [29], we set the circuit to close in 600000 seconds to ensure each batch was completed before closing. We also set *UseEntryGuards* to 0 to disable the set of limited entry guards and allow for switching of the entry guard. It is important to note that the client does not generally make these changes, and these modifications were made to gather realistic data and eliminate unrealistic advantages for the attacker [29].

2. Packet Capture (PCAP) Parser

Wireshark, a commonly used network protocol analyzer, was used in conjunction with the video crawler to collect Packet Capture (PCAP) files to be used for analysis. The PCAP size limit that was configured in the video crawler was up to ~120 Megabytes per video [32]. Image screenshots are also captured periodically during collection so that the adversary can visually verify that data from the YouTube videos is being collected as shown in Figure 3.

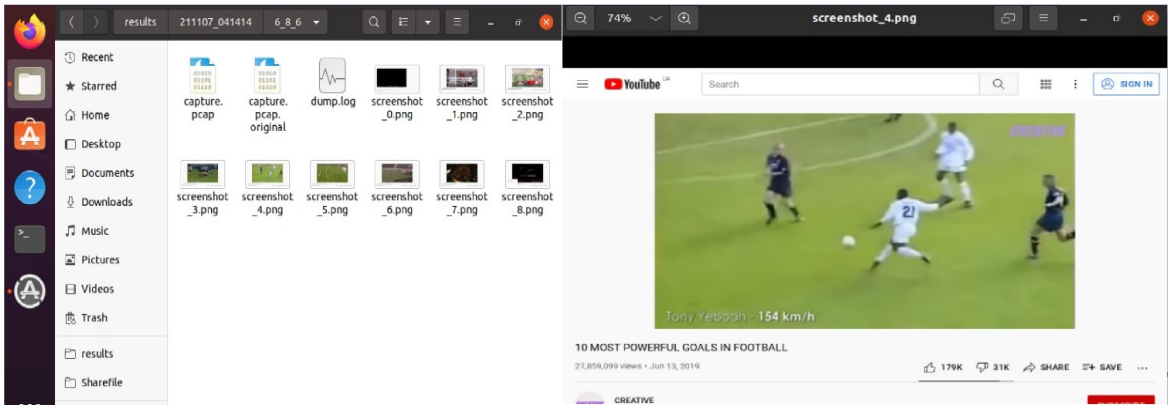

Figure 3. Screenshot capture on video crawler

A parser program was used to separate the collected data to analyze its traffic more efficiently. Sirinam et al. [4]. mention that "the attacker is assumed to be able to parse all traffic…and isolate it from other traffic." We used *Scapy*, a Python program interpreter, to first dissect the captures. Next, TShark (Wireshark's terminal-oriented version) was then used to filter the traces by IP/MAC address. The parser program will then process IP-level capture PCAP files into a sequence of tuples containing packet time and direction, and recursively search through directories to find the PCAP files [8]. Figure 4 illustrates the time and direction output of a traffic capture after being fully parsed (left column is packet time and right column is packet direction).

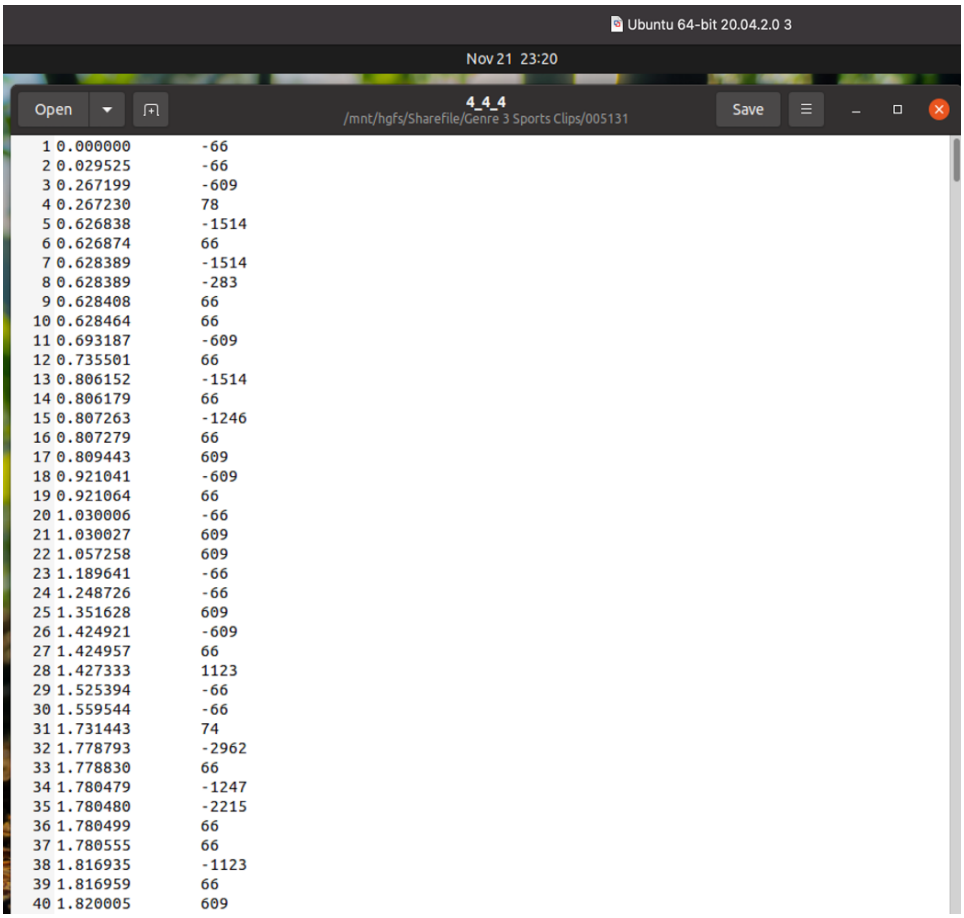

Figure 4. PCAP parser program output

Applying Wang and Goldberg's approach [29], we represent each traffic instance as a sequence of positive and negative integers. From the client's perspective, we later convert raw packet direction to positive (+1) for outgoing and negative (-1) for incoming to reduce and simplify the raw traffic traces as this allows for optimal performance of our DF models as demonstrated by Sirinam et al. [4].

The parser program was also useful for removing flawed instances during data collection. Failed connections to the server or server-oriented messages that denied access to YouTube would result in the video crawler aborting its operation resulting in an incomplete video dataset batch. These batches were deleted and not used for analysis. Following Wang and Goldberg [29], "if the size of the video traffic instance was less than

20% of the median size for that video, it was removed." In addition, samples that contained less than 100 packets of time and direction data were discarded.

3. Dataset Computation and Storing

Our research utilized computing and storage resources from the Naval Postgraduate School (NPS) Hamming supercomputer [33]. For our data processing, this allowed the use for one NVIDIA Quadro RTX 6000 graphics processing unit (GPU) with 24 gigabytes of memory, 4,608 CUDA parallel-processing cores, and 576 tensor cores [34].

Using Wang and Goldberg's methodology [29], "data was collected batch-by-batch with each corresponding to one circuit (one client)." Each batch of 90 videos took approximately 5–6 hours to download depending on home Wi-Fi network connectivity. The complete collection of one batch of videos with corresponding crawler log is shown in Figure 5.

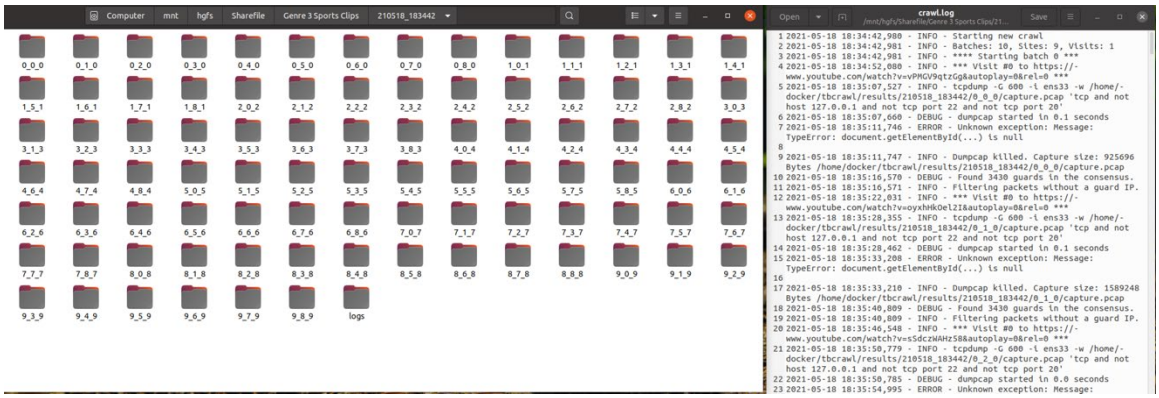

Figure 5. Video batch collection with crawler log

Following Campuzano's research [8], we added to his previous Disney movie video data collection of 138 batches for a final total of 165 batches. During our research, we also added two extra genres totaling 165 batches each to ensure equal amounts of data across all three video genres. Data collection took place over the course of approximately eight months. In order to nearly triple the dataset size from prior work [8], two computers were required to run the video crawler and PCAP parser programs. Upon collecting an additional

27 batches to add to the first genre of videos utilizing both computers, one computer was then designated to collect all data for the second genre and the other computer collected all data for the third genre to ensure equal dataset size throughout collection.

Our research is focused on the closed-world environment where the video ID and video genre are used as class labels in the training and test set in a manner similar to that used by Hermann et al. [23]. who used webpages as labels in their WF traffic analysis attack model. Using Hamming resources, the datasets were created and stored in a Pandas Data frame using Jupyter Notebook. The dataset was formatted into rows and columns; the rows corresponded to each video download, and the columns corresponded to video ID, video length, last packet time, individual packet times $\{t_0, t_1, t_2, \ldots t_n\}$, and individual packet directions $\{d_0, d_1, d_2, \ldots d_n\}$. Within the data frame, we decided to exclude packet times and lengths and only use packet directions as features for training the DF model based on Sirinam et al.'s [4] findings which showed that using other features such as packet timestamps do not provide a noticeable improvement in attack accuracy. In addition, according to Hermann et al.'s [23] research, relying on packet length frequencies was found to be rather unproductive and produced significantly lower accuracy results in their fingerprinting attacks. The total number of samples for our research was 19,831 video downloads. Figure 6 is a visual representation of the data collection framework for our research.

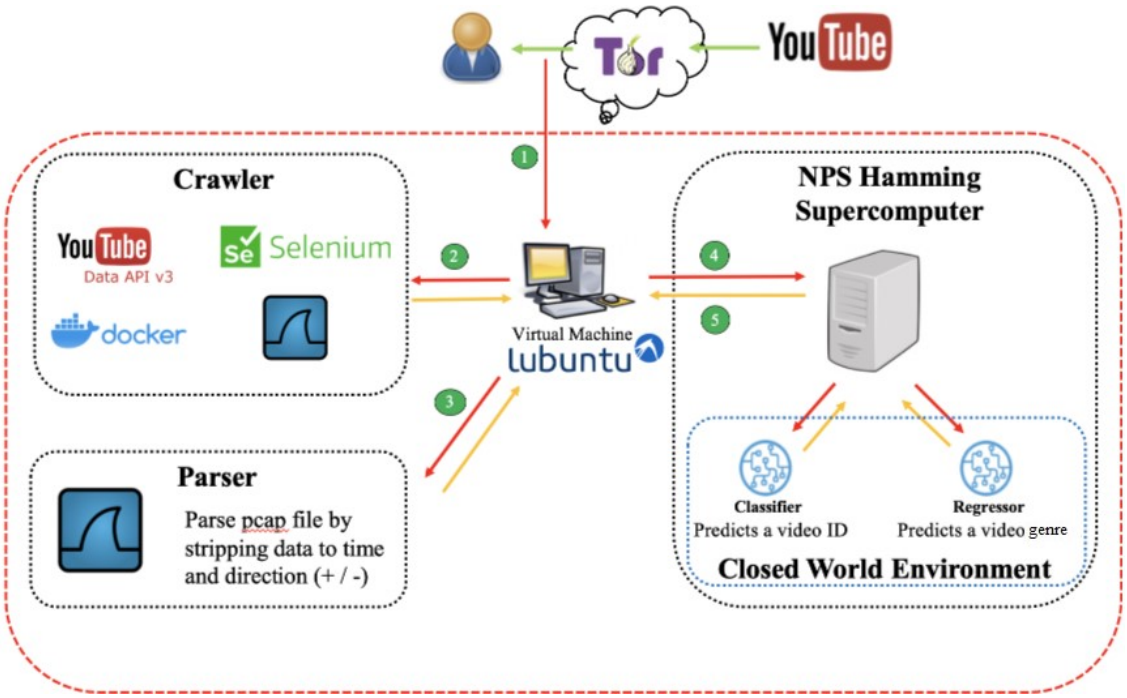

Figure 6. Data collection environment. Adapted from [8].

B. THREAT MODEL

Using the same concept as demonstrated in Sirinam et al. [4] and Campuzano's work [8], we assume that the adversary has the capability to compromise an internet service provider (ISP) or access the Wi-Fi network of the client to passively observe traffic. We, behaving as the adversary, passively observed and recorded traffic between the client and the first entry node to the Tor network. This approach included intercepting traffic traces without modifying or decrypting the transmission to predict information about the YouTube video the client is viewing over the Tor network [8].

Within our closed-world environment, we also assume that the client can only download a small set of YouTube videos such as the 27 videos used in our work for the adversary to gather samples and train and test on. As pointed out by Juarez et al.[22] a closed-world WF attack model does not resemble real-world characteristics because such would require an adversary having to consider a sample size comprised of every single website that exists on the internet; a sample size that would be so enormous and dynamic due to the constant flux of websites being created and taken down on a daily basis. In

addition, the client would have free reign to visit any website of his choosing. Similarly, an adversary with even the most capable and advanced resources would be unable to collect every single traffic trace for all videos available on YouTube. Our DF model simplifies the threat model and allows the adversary to wield the necessary capabilities to know in advance which videos the client could have streamed to illustrate how a small training set could still be useful as an initial approach to build a reliable dataset and identify recognizable patterns and deviations [8].

IV. TEST DESIGN AND IMPLEMENTATION

This chapter discusses the data analysis and implementation of the DF models in greater detail. This process includes adjusting CNN hyperparameters using Python DL modules, *Keras* and *Tensorflow*, to further explore Tor's lightweight defense models in comparison to the no defense VF attack scenario. Test results are examined and justified as we investigate the overall efficacy of defenses provided by DF in a VF attack closed-world scenario.

A. FEATURE SELECTION

Upon completion of collecting all batches for all three video genres, the dataset was scaled, and raw traffic packet directions were converted to +1 and -1 values. Positive (+1) values indicate outgoing packets while negative (-1) values indicate incoming packets with respect to the client. Based on Sirinam et al.'s [4] work, we selected only packet directions to be used as features as they provided the best performance for all three of the DF models. Figure 7 is the Pandas Data Frame that was used to compute and store the incoming and outgoing packets.

	d0	d1	d2	d3	d4	d5	d6	d7	d8	d9	...	c
row_1	-1.0	1.0	1.0	-1.0	1.0	-1.0	1.0	-1.0	1.0	-1.0	...	
row_7	-1.0	1.0	1.0	-1.0	-1.0	-1.0	-1.0	1.0	1.0	1.0	...	
row_10	-1.0	1.0	-1.0	-1.0	1.0	-1.0	-1.0	1.0	-1.0	-1.0	...	
row_11	-1.0	-1.0	1.0	1.0	1.0	-1.0	1.0	-1.0	1.0	-1.0	...	
row_12	-1.0	1.0	-1.0	-1.0	1.0	-1.0	-1.0	1.0	-1.0	1.0	...	
row_14	-1.0	1.0	-1.0	1.0	-1.0	1.0	1.0	-1.0	1.0	-1.0	...	

Figure 7. Scaled dataset representing only packet direction (+1 outgoing, -1 incoming)

Following Wang and Goldberg [29], we used packet direction as independent *x* values to be used as input to three different DF models presented by Sirinam et al. [4]:

WTF-PAD, W-T, and NoDef. The corresponding *y* values, otherwise known as classes, are the YouTube video IDs and genre type.

Using the *Scikit-learn* library [35] in Python, the *train_test_split* utility allowed us to divide up our data into train and test subsets. Amongst the total dataset of 19,381 video samples, the data was scaled and split into a train and test set where 70% of the data was used for training and 30% was used for testing. We also utilized a validation set that used 50% of the test set. The validation set is used to "provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters" [36]. In other words, validating our data allows us to compare models that have different hyperparameters and select the ones that achieve the highest accuracy on the validation set.

B. TRAINING

Once our dataset was scaled, traffic packet directions were converted to simple positive and negative integers and split into training, validation, and test sets. Python DL libraries *Keras* and *Tensorflow* were used in conjunction with Hamming resources to train and evaluate our DF models. *Keras* was used as the front-end library and *Tensorflow* was used as the back-end for our DF evaluations.

Training epochs were instrumental in achieving our results for our three DF attack models. *Training epochs,* otherwise known as iterations, refer to "the number of passes of the entire training dataset the machine learning algorithm has completed" [37]. Sirinam et al. [4]. utilized 30 *training epochs* in their research to achieve optimal training accuracy. Campuzano [8] in his work found that 60 *training epochs* achieved the highest accuracy levels. Initially, we implemented 60 *training epochs* to our dataset, but found that we did not achieve our highest accuracy score amongst all three DL models using this parameter. We proceeded to increase our *training epochs* to 120 which yielded the highest accuracy scores for all three models for both video ID and video genre predictions. As Sirinam et al. [4]. mentions, "the classifier gradually learns better with more training epochs".

C. TEST RESULTS

Our approach was to train the DF models with two classes of predictors: 1) to predict video ID, and 2) to predict video genre. For the closed-world setting, equal sample size and same number of *training epochs* were used for all three DF models to ensure consistency and validity of results.

1. Video ID

Video ID prediction for our DF models consisted of training and testing our classifiers on a total of 27 videos with each having its assigned unique numerical video ID ranging from 0–26 as shown in Table 1.

Table 1. YouTube video ID list

0	Frozen 2	**9**	Demi Lovato Dancing	**18**	Baseball: Best Catches
1	Toy Story 4	**10**	Lil Nas X: Montero	**19**	Tennis: AUS Open 2012
2	Dark Phoenix	**11**	Gwen Stefani: Slow Clap	**20**	NFL: Best Tackle Moments
3	The Lion King	**12**	Cardi B: Up	**21**	Ice-skating: Olympics 2016
4	Dora and the Lost City of Gold	**13**	Lil Nas X: Rodeo	**22**	Gymnastics: Olympics 2016
5	Rambo: Last Blood	**14**	Ariana Grande: Positions	**23**	Swim: Michael Phelps 200m
6	Aladdin	**15**	Kenia Os: Tu Peor Pesadilla	**24**	Highest Olympic Jumps
7	Joker	**16**	Polo G: Rapstar	**25**	Snowboarding: Shaun White
8	Terminator 6: Dark Fate	**17**	Machine Gun Kelly: Daywalk	**26**	Soccer: 10 Best Goals

Note. Video ID numbers are highlighted in yellow. Video titles are in blue-shaded cells.

Evaluations of each DF model took approximately 2–3 hours to complete 120 *training epochs* using one GPU from NPS Hamming resources. Without using a GPU, each

model would take approximately 24 hours to complete. Figure 8 is a graphical representation of our results for video ID prediction for all three DF models.

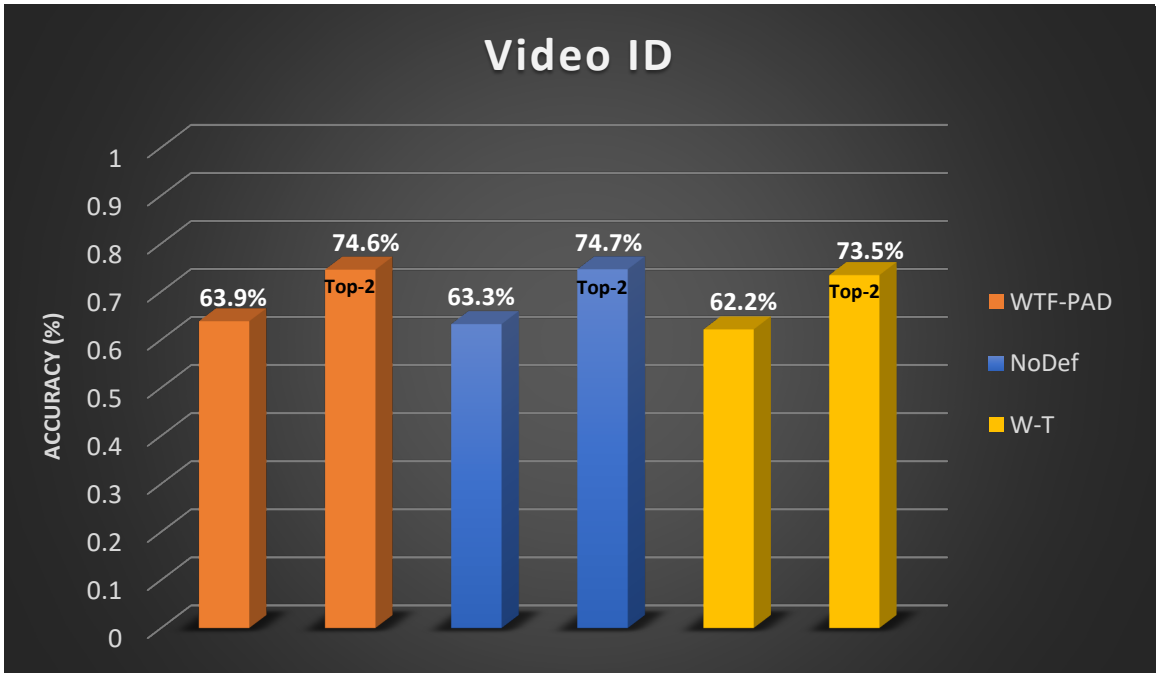

Figure 8. Attack accuracy on predicting YouTube video ID

For the defenseless NoDef model, the attack accuracy was 63.3%. For WTF-PAD and W-T defenses, the scores were 63.9% and 62.2% respectively. As for Top-2 accuracies, W-T had the lowest attack accuracy at 73.5% indicating that the true class matched with any two of the most probable classes in the predicted model. For WTF-PAD and NoDef, the Top-2 scores were 74.6% and 74.7% respectively. In Campuzano's [8] work, he achieved a score of 84% for NoDef. For WTF-PAD and W-T defenses, he scored 84% and 83%, respectively. Although his accuracy scores may appear relatively higher than ours, our research introduced a 27-class problem versus the 9-class problem used in his work. Therefore, our DF classifier performed strongly for our VF attack scenarios especially as the number of video ID classes used in our research was three times larger than that in [8].

2. Video Genre

Video genre prediction for our DF models consisted of training and testing our classifiers in the same manner as video ID except for using video genre type as the class label instead of video ID. As shown in Table 2, the three video genre classes are labeled 0, 1, and 2 for Disney Movie Trailers, Music Videos, and Sports Clips, respectively.

Table 2. YouTube video genre list

Genre 0 Disney Movie Trailers		**Genre 1 Music Videos**		**Genre 2 Sports Clips**	
0	Frozen 2	**9**	Demi Lovato Dancing	**18**	Baseball: Best Catches
1	Toy Story 4	**10**	Lil Nas X: Montero	**19**	Tennis: AUS Open 2012
2	Dark Phoenix	**11**	Gwen Stefani: Slow Clap	**20**	NFL: Best Tackle Moments
3	The Lion King	**12**	Cardi B: Up	**21**	Ice-skating: Olympics 2016
4	Dora and the Lost City of Gold	**13**	Lil Nas X: Rodeo	**22**	Gymnastics: Olympics 2016
5	Rambo: Last Blood	**14**	Ariana Grande: Positions	**23**	Swim: Michael Phelps 200m
6	Aladdin	**15**	Kenia Os: Tu Peor Pesadilla	**24**	Highest Olympic Jumps
7	Joker	**16**	Polo G: Rapstar	**25**	Snowboarding: Shaun White
8	Terminator 6: Dark Fate	**17**	Machine Gun Kelly: Daywalk	**26**	Soccer: 10 Best Goals

Note. Video genres are highlighted in yellow. Video ID and titles are in blue-shaded cells.

Evaluation for each DF model took roughly the same amount of time to complete 120 *training epochs* as video ID prediction. The highest level of accuracy was also achieved using 120 *training epochs*. Figure 9 is a graphical representation of our results for video genre prediction for all three DF models.

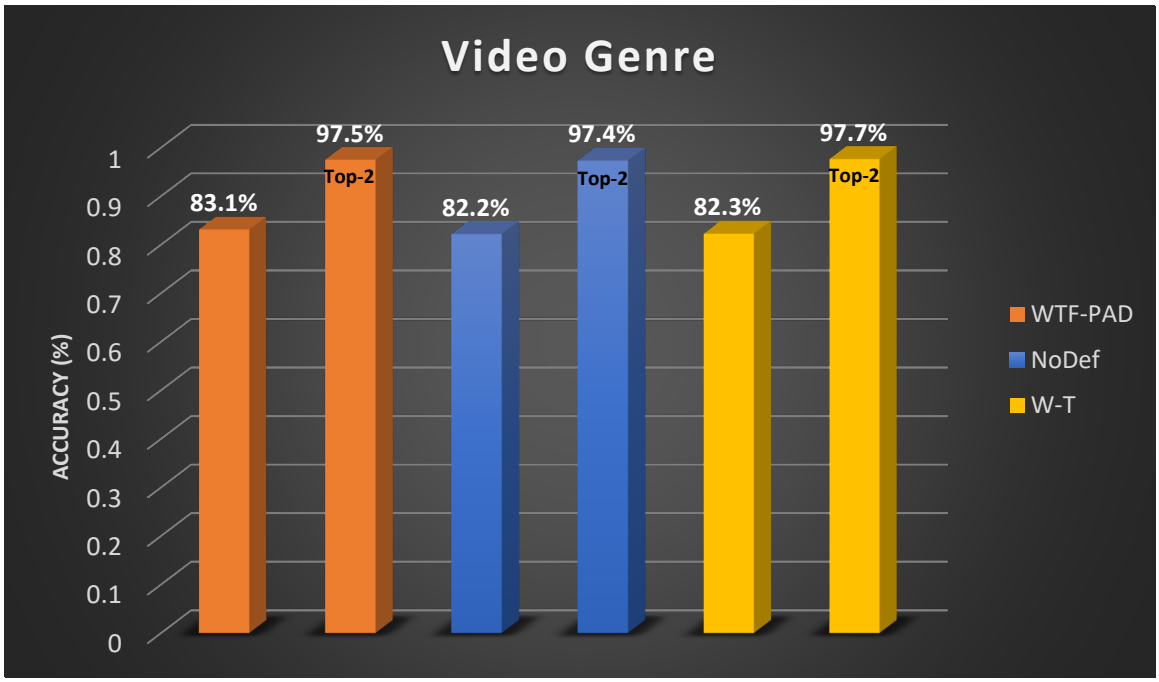

Figure 9. Attack accuracy on predicting YouTube video genre

Overall, video genre prediction was successful in achieving significantly higher accuracy scores than those from the video ID attacks. For the defenseless NoDef model, the attack accuracy was 82.2% with a Top-2 accuracy of 97.4%. For WTF-PAD, the attack accuracy was 83.1% with a Top-2 accuracy of 97.5%. Lastly, W-T achieved an accuracy score of 82.2% as well as the highest Top-2 score of 97.7%. When comparing the video ID attack versus video genre attack, the higher scores were somewhat expected as the number of classes reduced from 27 (video IDs) to 3 (genres) for the models to train and test on.

Since Campuzano's [8] research consisted of only one genre type of YouTube videos and focused only on video ID prediction, we could not do a direct comparison of his results with ours for this particular attack scenario. However, it is interesting to see how our results for the VF attack with video genre type predictions closely resemble his results, and in fact, achieved higher accuracy for two out of the three models: W-T and WTF-PAD.

3. Experimental Findings

Once all DF models have been trained and evaluated, we are able to analyze the significance of our results for both Tor users and the adversaries who execute similar VF

attacks. For both video ID and genre prediction, our VF attacks were quite successful by achieving approximately 64% and 83% accuracy, respectively. Our results also show that neither WTF-PAD nor W-T are effective defenses as they do not appear to influence accuracy scores significantly or consistently.

Such outcomes may appear shocking to an average Tor user who believes he achieves complete privacy and anonymization using Tor services. Moreover, adversaries are taking advantage of these loopholes that exist on Tor to degrade overall security on the network by intruding into user privacy and developing more advanced techniques to reveal video content being streamed online by unknowing users. We conclude that all three DF models exemplify the same important concept of privacy concern for Tor users and the high effectiveness of VF attacks.

Our research yielded results that showed that greater attack accuracy was surprisingly achieved *with* defenses added for our DF VF attacks. This is counterintuitive because clients reinforced with defenses should theoretically make fingerprinting attacks more challenging for the adversary thereby resulting in lower attack accuracy. For both video ID and video genre classification, the VF attack against the NoDef client resulted in the lowest accuracy scores thereby revealing the least amount of information to the attacker.[1] For predicting video genre, the W-T defense client's Top-2 accuracy resulted in the highest score at 97.7%. Campuzano [8] observed similar results indicating that the lightweight defenses did not provide a significant reduction in accuracy compared to NoDef, though his work utilized only one genre of videos consisting of 138 video batches and 3,810 training samples. His NoDef client achieved the highest score of 84% accuracy, and 82% and 81% for WTF-PAD and W-T, respectively [8] for video ID prediction.

A possible explanation for these results may be related to sample size gathered in our research. According to Bhat et al. [38], "one significant drawback of deep learning, however, is that it generally requires a large amount of training data…performance issues in low-data scenarios can be a serious issue for WF attacks: since website traces change

[1] With the exception of NoDef vs. W-T for video ID prediction in which NoDef scored 0.012% higher – a relatively negligible margin.

quickly." In Sirinam et al.'s [4] work, they collected 1,000 traces for each of 95 websites; a dataset that was considerably larger than ours. Similar to website traces, there is a possibility for a significant amount of variability of video traces which is especially common with YouTube when presented with server-error messages or CAPTCHA pop-ups that require user intervention to resolve challenge-response authentication.

Another possible explanation could be that the adaptive padding and dummy packets used for WTF-PAD and W-T defenses are less effective on traffic bursts from video traffic traces in comparison to website traces. As Sirinam et al. [4] mentions, the adaptive padding algorithm is used to make time gaps between traffic bursts less apparent thereby making it more difficult for the adversary's classifier to identify features of the websites. In [7], Schuster et al. discusses how "video streams are known to be bursty" due to its "initial short period of buffering followed by the steady state of alternating 'On' (short bursts of packets) and 'Off' periods." Assuming our training data was sufficient for our DF algorithms, this could potentially point to the idea that adding extra padding or dummy packets interestingly aids the attacker's classifier in extracting features from video traces to make accurate predictions.

Lastly, the DF models Sirinam et al. [4]. used were "specifically designed to effectively perform WF attacks by leveraging the state-of-the-art techniques from computer vision research." Our research closely resembled Sirinam et al.'s [4] work in terms of using the same DF models, hyperparameters, dataset format, and selection of features. There is a strong possibility that greater modifications were required to better fit our VF attack scenarios since they involved video streaming vice webpages along with evaluating accuracy on different class labels.

V. CONCLUSION AND FUTURE WORK

This research expanded upon previous work involving VF attacks using DL models over Tor. Specifically, our VF attack was constructed using a new fingerprinting attack method called DF proposed by Sirinam and Juarez et al. [4], [22] for use in a closed-world setting. DF utilizes CNN architectural design to determine which videos a Tor user may be streaming over an encrypted connection based on selected features from raw traffic captures.

A. RECOMMENDATIONS FOR FUTURE WORK

We believe that the validity and effectiveness of these DF models depend on understanding the appropriate hyperparameters, sample size, and network configurations that are required to produce reliable, consistent results. Further research may be conducted to train and test on much larger sample sizes that include a greater variety of video types, lengths, and streaming sources. Following Campuzano's [8] work, we successfully added two more video genres to evaluate how well the classifier can predict video genre in addition to video ID. One idea would be to continue adding more video batches and genres and include videos from other video streaming sources such as Dailymotion and Vimeo that are growing in popularity around the world.

As Campuzano [8] and our research have shown, the defense models proposed by Sirinam and Juarez et al. [4], [22] did not perform as well for our VF-designed attacks. Our research showed that attack accuracy *increased* with the use of defensive measures. Sirinam and Juarez et al.'s [4], [22] research (which involved webpages and WF attacks) showed the opposite, that the defenses *decreased* attack accuracy. These contradicting results indicate that further work dedicated to VF attack scenarios using defense models is needed. Supplemental research into how dummy packets, low latency overhead structures, and operating in half duplex modes could be beneficial to discovering possible adverse effects on these defense models thereby making it easier for an adversary to conduct more effective VF attacks. It would be interesting to see whether the results we obtained are replicated in a WF attack scenario. Conversely, one may attempt to further improve

defenses, such as WTF-PAD and W-T, against VF attacks to make it more difficult for an attacker to obtain higher video prediction accuracy. This could prove to be invaluable to those on the user-side who wish to better protect their privacy and enhance defensive measures against attackers.

Finally, future work including an open-world scenario could provide a more realistic environment that would involve the attacker *not* being able to download all the videos the user has downloaded. Thus, the attacker must increase his collection of traces by a substantially larger amount than what is collected in a closed-world scenario. In addition, the attacker must select particular class labels that are better suited for training and testing the DF models in an open-world setting. By acquiring and using more computers and other capable network resources, larger sample sizes may be obtained thereby optimizing DF model performance and results.

B. CONCLUSION

The main benefit that most seek from using Tor is to remain anonymous online. As our research and that of several others has shown, Tor is vulnerable to fingerprinting attacks that unmask the digital content that a user is visiting or viewing while an adversary passively observes communications between the user and entry node. Raw traffic packets and patterns can surprisingly reveal enough data and information that allow an outsider to make accurate attack predictions and circumvent the protective layers of an encrypted channel.

For regular Tor users, our research may serve as a reminder that using anonymization services cannot fully conceal one's online activities. Attackers are developing new and more advanced attack vectors that threaten online user privacy. These may be deployed in the network, transport, and application layers. Not to mention, how common it is for an average person to connect to insecure Wi-Fi networks and for ISPs to be compromised thereby heightening users' vulnerability. To address these problems, it is our hope that our findings and techniques may be used to formulate more advanced means to understand the distribution of video material via the Dark Web in order to prevent the spread of terrorism and malicious activities.

Although our research solely focused on a controlled closed-world environment where fewer dynamic variables are considered, one should not underestimate the growing complexity and effectiveness of attacks that are being continuously developed, and the lack of additional defensive measures a typical user could initiate on his own behalf when using Tor services.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

[1] A. Lewman, "Journalists use tor to communicate more safely with whistleblowers and dissidents. Nongovernmental organizations (NGOs) use tor to allow," *Advances in Cyber Security: Technology, Operation, and Experiences*, pp. 109, 2013.

[2] T. Sabbah, A. Selamat, H. Selamat, R. Ibrahim, and H. Fujita, "Hybridized term-weighting method for dark web classification," *Neurocomputing*, *173*, pp. 1–38, 2016. [Online]. Available: https://www.researchgate.net/profile/Thabit-Sabbah/publication/282586721_Hybridized_Term-Weighting_Method_for_Dark_Web_Classification/links/5c62725f92851c48a9cd56c5/Hybridized-Term-Weighting-Method-for-Dark-Web-Classification.pdf

[3] A. Salem, E. Reid, and H. Chen, "Multimedia content coding and analysis: Unraveling the content of Jihadi Extremist Groups' videos," *Conflict and Terrorist, 31:7*, 2008, pp. 605–626.

[4] P. Sirinam, M. Imani, M. Juarez, and M. Wright, "Deep fingerprinting: Undermining website fingerprinting with deep learning," *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 1928–1943.

[5] A. Lashkari, G. Draper-Gil, M.S. Mamun, and A. A Ghorbari, "Characterization of tor traffic using time based features," *ICISSp*, 2017, pp. 253–262.

[6] J. Lu, "Video fingerprinting for copy identification: from research to industry applications," *Media Forensics and Security Volume 7254*, 2009, pp. 725402.

[7] R. Schuster, V. Shmatikov, and E. Tromer, "Beauty and the burst: Remote identification of encrypted video streams," *26th {USENIX} Security Symposium {USENIX} Security 17)*, 2017, pp.1357-1374.

[8] C. Campuzano, "Towards video fingerprinting attacks over Tor," M.S. thesis, Naval Postgraduate School, Sept. 2021. [Online]. Available: https://calhoun.nps.edu/handle/10945/68304

[9] R.W. Gehl, "Weaving the dark web: legitimacy on freenet, Tor, and I2P," MIT Press, Aug 2018. [Online]. Available: https://www.google.com/books/edition/Weaving_the_Dark_Web/RzdmDwAAQBAJ?hl=en&gbpv=1&dq=the+dark+web+design+anonymity&pg=PR7&printsec=frontcover

[10] P. Biddle, P. England, M. Peinado, and B. William, "The darknet and the future of content protection," *ACM Workshop on Digital Rights Management*, 2-2, pp. 155–176.

[11] S. Samtani, H. Zhu, and H. Chen, "Proactively identifying merging hacker threats from the Dark Web: A diachronic graph embedding framework (D-GEF)," *ACM Transactions on Privacy and Security Volume 23, Issue 24*, 2020, pp. 1–33.

[12] R. Ehney and J. Shorter, "Deep Web, Dark Web, Invisible Web and the Post ISIS World," *Issues in Information Systems Volume 17, Issue IV*, 2016, pp. 36–41.

[13] R. Scrivens, G. Davies, R. Frank, "Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors," *Behavioral sciences of terrorism and political aggression*, 2018, pp. 39–59.

[14] M. Schafer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "BlackWidow: Monitoring the dark web for cyber security information," *2019 11th International Conference on Cyber Conflict (CyCon)*, 2019, pp. 1–21.

[15] K. Swan, "Onion Routing and Tor," *Georgetown Law Tech Review*, 2016, pp. 110–118.

[16] Y. Sun, A. Edmundson, L. Vanbever, O. Li, J. Rexford, M. Chiang, and P. Mittal, "{RAPTOR}: Routing attacks on privacy in tor," *24th {USENIX} Security Symposium ({USENIX} Security 15*, 2015, pp. 271–286.

[17] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," *Proceedings of the 13th Conference on USENIX Security Symposium*, USENIX Association, 2004.

[18] R. A. Haraty and B. Zantout, "The TOR data communication system: A survey," *2014 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2014, pp. 1–6.

[19] V. Lapshichyov, "TLS certificates of the Tor network and their distinctive features," *International Journal of Systems and Software Security and Protection Volume 10, Issue 2*, 2019, pp.20-43.

[20] Y. Gilad and A. Herzberg, "Spying in the dark: TCP and Tor traffic Analysis," *International symposium on privacy enhancing technologies symposium*, 2012, pp. 100–119.

[21] W. Cai, G Gou, P. Fu, M. Jiang, Z. Li, and G. Xiong, "JumpEstimate: a Novel Black-box Countermeasure to website Fingerprint Attack Based on Decision-boundary Confusion," *2020 IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1–7.

[22] M. Juarez, S. Afroz, G. Acar, C. Dias, and R. Greenstadt, "A critical evaluation of website fingerprinting attacks," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pp. 263–274.

[23] D. Hermann, R. Wendolsky, and H. Federrath, “Website fingerprinting: attacking popular privacy enhaving technologies with the multinomial naïve-bayes classifier,” *Proceedings of the 2009 ACM workshop on Cloud computing security*, 2009, pp. 31–42.

[24] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel, “Website fingerprinting in onion routing based anonymization networks,” *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, 2011, pp. 103–114.

[25] A. Rao, Y. Lim, C. Barakat, A. Legout, D. Towsley, and W. Dabbous, “Network characteristics of video streaming traffic,” *Proceedings of the Seventh Conference on emerging Network Experiments and Technologies*, 2011, pp. 1–12.

[26] Sandevine. (2018, Oct). The Global Internet Phenomena Report October 2018. [Online]. Available: https://www.sandvine.com/hubfs/downloads/phenomena/2018-phenomena-report.pdf

[27] Y. Li, Y. Huang, R. Xu, S. Seneviratne, K. Thilakarathna, A. Cheng, D. Webb, and G. Jourjon, “Deep Content: Unveiling video streaming content from encrypted WiFi traffic,” *2018 IEEE 17th International Symposium on Networking Computing and Applications (NCA)*,” 2018, pp. 1–8.

[28] V. Rimmer, D. Preuveneers, M. Juarez, T. Van Goethem, and W. Joosen, “Automated website fingerprinting through deep learning,” *arXiv preprint arXiv:1708.06376*, 2017.

[29] T. Wang and I. Goldberg, “Walkie-talkie: An efficient defense against passive website fingerprinting attacks,” *26th {USENIX} Security Symposium {USENIX} Security 17)*, 2017, pp. 1375–1390.

[30] V. Shmatikov and M.H. Wang, “Timing analysis in low-latency mix networks: Attacks and defenses,” *European Symposium on Research in Computer Security*, 2006, pp. 18–33.

[31] T. Wang and I. Goldberg, “On realistically attacking Tor with website fingerprinting,” *Proc. Priv. Enhancing Technol.*, 2016, pp. 21–36.

[32] N. Mathews. (2019, Jun). Tor-browser-crawler-video. [Online]. Available: https://github.com/notem/tor-browser-crawler-video/

[33] Naval Postgraduate School, “Available HPC Resources,” *Naval Postgraduate School*. [Online]. Available: https://nps.edu/web/technology/hpc

[34] Nvidia. Nvidia Quadro RTX 6000 Graphics Card. [Online]. Available: https://www.nvidia.com/en-us/design-visualization/quadro/rtx-6000/

[35] D. Cournapeau. (2007). Sklearn. [Online]. Available: https://scikit-learn.org/stable/

[36] J. Brownlee. (2020). What is the difference between test and validation datasets? [Online]. Available: https://machinelearningmastery.com/difference-test-validation-datasets/

[37] D. Bell. (2020). Epoch (machine learning). [Online]. Available: https://radiopaedia.org/articles/epoch-machine-learning?lang=us

[38] S. Bhat, D. Lu, A. H. Kwon, and S. Devadas, "Var-cnn: A data-efficient website fingerprinting attack based on deep learning," *Proceedings on Privacy Enhancing Technologies*, pp. 292–310, 2019. [Online]. Available: https://sciendo.com/downloadpdf/journals/popets/2019/4/article-p292.pdf

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California