7-2023

Using Deep Learning for Encrypted Traffic Analysis of Amazon Echo

Surendra Pathak
The University of Texas Rio Grande Valley

USING DEEP LEARNING FOR ENCRYPTED TRAFFIC ANALYSIS OF AMAZON ECHO

A Thesis

by

SURENDRA PATHAK

Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Major Subject: Computer Science

The University of Texas Rio Grande Valley

July 2023

USING DEEP LEARNING FOR ENCRYPTED TRAFFIC ANALYSIS OF AMAZON ECHO

A Thesis
by
SURENDRA PATHAK

COMMITTEE MEMBERS

Dr. Emmett Tomai
Chair of Committee

Dr. Bin Fu
Committee Member

Dr. Dong-Chul Kim
Committee Member

July 2023

ABSTRACT

Pathak, Surendra, Using Deep Learning for Encrypted Traffic Analysis of Amazon Echo . Master of Science (MS), July, 2023, 44 pp., 6 tables, 12 figures, references, 34 titles.

The adoption of the Amazon Echo family of devices in modern homes has become very widespread at the current time, with hundreds of millions of devices sold. Moreover, the global smart speaker market size is growing vigorously and is projected to continue to bigger. Smart speakers allow users hands-free interaction by allowing voice control, promoting human-computer interaction to greater avenues. Though smart speaker can be useful assistant, it has some serious security concerns that need to be studied.

In this study, an analysis of the security and privacy concerns of smart speakers is presented along with a passive attack, namely voice command fingerprinting. We start by introducing different security vulnerabilities of Amazon Alexa. Then, a voice command fingerprinting attack is implemented. In a voice command fingerprinting attack, an attacker eavesdropping on encrypted communication traffic can infer users' voice commands. The attacker can use side channel information like packet length, direction, and order of traffic between Amazon Echo and the cloud server to make predictions of voice commands issued by the user. Different ensemble strategies are implemented to increase attack performance. Stacked generalization has a superior performance among all attacks, correctly predicting 90.54% of voice commands. The details on implementation techniques and experimental evaluation are also presented in this work.

DEDICATION

I would like to dedicate this to my parents Mr. Tuk Narayan Pathak, and Mrs. Chandra Maya Pathak, who put unwavering confidence in me and were with me in every single step of this journey. It would not be possible without your love and support.

Also, I would like to extend my gratitude to family, friends, and professors who were a part of this journey.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to all the people that provided support for the completion of my Master's thesis. Firstly, I would like to thank both of my present and previous advisors, Dr. Emmett Tomai and Dr. Sheikh Ariful Islam respectively. Both your mentorship and guidance were invaluable in shaping the direction of my research work and getting the desired results timely.

Secondly, I would like to thank Dr. Bin Fu and Dr. Dong-Chul Kim for serving on my thesis committee and providing valuable feedback.

I extend my sincere appreciation to the College of Engineering and Computer Science for providing me an opportunity to serve as a Research Assistant under Presidential Research Fellowship program. The opportunity provided me with a platform to develop myself as a researcher and succeed as a graduate student.

Finally, I would like to thank my parents, Mr. Tuk Narayan Pathak and Mrs. Chandra Maya Pathak, my siblings Bimala, Srijana, and Rabindra for their continuous love, support, and encouragement. I also would like to thank all the friends who were part of this journey.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

A smart speaker is an intelligent voice-activated loudspeaker device that has virtual assistant software, which is capable of performing various tasks and providing information or services. Usually, it includes elements like an audio output speaker and an integrated virtual assistant, such as Amazon Alexa, Google Assistant, or Apple Siri. An instance of interaction between a user and a smart speaker might include the user telling "Alexa, give me today's weather forecast" to which the device could respond by playing a summary of the day's weather forecast. Smart speakers primarily provide a wide range of functionalities, which includes tasks like playing music, answering queries, controlling smart home devices, and various other activities. They can also be used as central control hubs for smart homes, allowing users to control smart home devices with voice commands.

1.1 Motivation

The widespread adoption of the Amazon Echo family of devices has made Intelligent Virtual Assistant (IVA) ubiquitous in modern homes. More than 100 million devices have been sold by January 2019 that have Alexa on board [6]. Similarly, the global smart speaker market size is growing tremendously and can reach a worth of USD 15.6 billion by 2025.

The device's popularity is partially attributed to its ability to carry out tasks using voice commands, which promotes human-computer interaction to a higher stage and abandons touch-based or other physical interaction-based interfaces. Though the new avenue of interaction has transcended device usability, it also introduces unforeseen security concerns. In 2017, a broadcast event triggered Amazon Echo in multiple households while covering an incident related to Ama-

zon Echo [22]. Malicious skills that have similar names to genuine skills can be created to collect user information [19]. Additionally, inaudible voice commands can be used to exploit Alexa and carry out attacks [33]. In addition, since Intelligent Virtual Assistants (IVAs) are very intrusive to users' personal space, proper security and privacy concerns must be assessed. Thus, these systems become more prone to attacks without proper research and analysis of underlying security vulnerabilities and privacy concerns. In that regard, different types of software, hardware, and system vulnerabilities are systematically studied and presented to provide a background. Then, a specific attack, namely Voice Command Fingerprinting, is implemented, and the result evaluation is presented.

1.2 Contibutions

A passive attack, namely voice command (VC) fingerprinting, of encrypted traffic data of Amazon Echo is proposed in this thesis. The attack was carried out employing multiple deep-learning algorithms. These algorithms predicted the voice command issued by users to their echo devices. The contribution to this work is listed below:

- Extensive study of security and privacy issues of Amazon Echo is presented in this work. The vulnerabilities are categorized into software, hardware, and system vulnerability for a systematic study. Similarly, corresponding mitigation techniques for vulnerabilities are also presented.
- Existing work on VC fingerprinting of Wang *et al.* [31] is established and is further extended with evidence. The weighted average ensemble technique was further investigated with different combinations of base models. In addition to the ensemble model with three base models presented in [31], we also implemented three different ensemble models that contain two base models each. We concluded that the model with CNN and SAE as base models performs the best among all four ensemble models.
- A more advanced ensemble technique, namely stacking or stacked generalization, is implemented. Three base estimators are trained on the input data, then a combiner final estimator algorithm is trained on the cross-validated predictions of base estimators to make a final

prediction. The three deep learning algorithms are used as base estimators, and Logistic Regression is used as a final estimator in the study. We have established that the stacked generalization provides superior performance than the existing weighted average ensemble technique mentioned above. Similarly, different combinations of base models are studied for stacked generalization.

- The findings of the VC fingerprinting attack highlight the necessity of a robust defense mechanism to be implemented by smart speaker manufacturers in order to prevent user data leakage.

1.3 Outline

In Chapter II, terminologies and experimental techniques are introduced along with the Amazon Alexa ecosystem. Then, the findings of the literature review on security issues of Amazon Echo are presented in Chapter III. The vulnerabilities are categorized into three subcategories, software, hardware, and system vulnerability, for a systematic study. In Chapter IV, detailed information on experimental methods is presented. The experimental results and evaluation, along with the dataset information, are presented in Chapter V. The conclusion of the study, along with possible future works, is presented in Chapter VI.

CHAPTER II

BACKGROUND

2.1 Terminology

2.1.1 Skill

Skills are the voice-driven capabilities developed for Alexa to power Amazon devices, such as Echo [1]. Alexa skills allow users to engage and enhance the functionality of Alexa to accomplish a diverse array of activities, including playing music, checking weather conditions, managing smart home devices, receiving news updates, making purchases, and utilizing various other services. Amazon skills are developed using *Alexa Skills Kit*, which is a set of APIs, tools, and documentation provided by Amazon.

2.1.2 Traffic Traces

A traffic trace encompasses comprehensive data about individual packets, such as their source and destination IP addresses, port numbers, protocol type (like TCP or UDP), packet size, timestamp, and data payload. In the case of smart speakers, traffic trace pertains to the sequence of network traffic packets linked to a user's command directed at the smart speaker and the subsequent response from the service provider (SP). In traffic analysis attacks, traffic traces are captured and saved as pcap files; the files may be converted to other formats, such as xls, before analysis.

2.1.3 Fingerprinting Attack

A Fingerprinting attack is primarily a penetration technique to gather a system's configuration information. The technique may involve scanning device network traffic or sending custom packets toward the device. A fingerprinting attack gathers different details about a user's system, in-

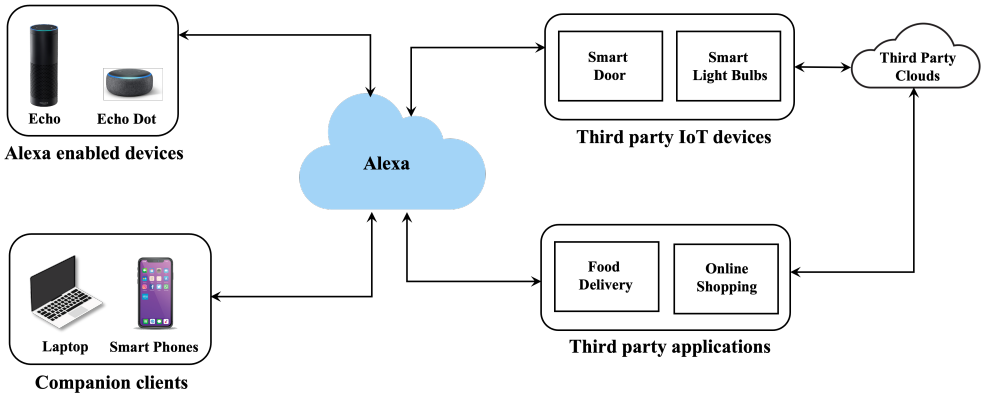

Figure 2.1: Amazon Alexa Ecosystem

cluding browser version, operating system, screen resolution, installed plugins, and other attributes. The information collected is exploited to identify vulnerabilities for a successful attack.

2.1.4 Accuracy

Accuracy is a metric used to evaluate the performance of attacks proposed in the study [18]. An attack is successful when an adversary can correctly identify the label of unlabelled traffic. If an adversary can correctly identify α unlabelled traces out of N total unlabelled traces, accuracy is presented as:

$$Accuracy = \frac{\alpha}{N}$$

Accuracy is a standard metric used in literature to compare the performance of Fingerprinting attacks.

2.2 Amazon Alexa Ecosystem

Amazon Echo has become a very popular virtual assistant in the past few years. The services offered by the device have benefited many households and even businesses. Using the Amazon Echo, users can easily carry out multiple actions by just speaking to the device, a futuristic living experience that was thought of as fiction a few years ago. Though many people may not be aware of it, there is a solid architecture to carry out these functionalities. There are several entities playing roles in this ecosystem. Each of the entities shown in Figure 2.1 is discussed in this section.

2.2.1 Alexa-enabled devices

Alexa-enabled devices are the Amazon Echo family of devices that a user interacts with, usually by speaking out a command. The device consists primarily of a microphone and speaker and is connected to the Internet. A wake word is used to activate the device. After activation, it starts recording voice which is passed to the Alexa voice service, where computation is done. When the computation is complete, it receives a response that is played as sound.

2.2.2 Alexa cloud services

Most of the computation of Intelligent Voice Assistant is carried out in Alexa cloud services. The voice commands are sent to Amazon Echo, and the response is stored in Alexa cloud services. Alexa cloud service composes entities that carry out Automatic Speech Recognition, Speech-Language Understanding, Natural Language Understanding, Text-to-Speech conversion, etc.

2.2.3 Companion clients

Devices running one of the Alexa companion applications, such as Amazon Alexa, are companion clients. Apart from interacting with Alexa using voice commands, users can interact with them through a companion app. Though there is no specific companion application native to personal computers, users can still access Alexa using the web browser from a personal computer.

2.2.4 Third-party Internet of Things (IoT) devices

Compatible IoT devices increase the usability of Amazon Echo by adding additional voice-controlled functionalities. With the growing adaptation of Amazon Echo, the number of compatible IoT devices is also increasing. Some of the popular compatible IoT devices include Philips Hue, Lifx Mini, August WiFi Smart Lock, etc.

2.2.5 Third-party applications

The functionality of Amazon Echo is enhanced by many third-party applications that extend Alexa's capabilities. In addition, the "skills" extend Alexa's functionality, enriching user experi-

ence and enabling user-tailored services. Some examples are *Lyft* (ride-sharing), *Domino's* (food ordering), *The Wall Street Journal* (news updates), etc.

2.3 Voice Command Fingerprinting

Voice Command Fingerprinting is a novel attack that is a passive attack, where an attacker eavesdropping on encrypted communication traffic can infer users' voice commands. Though encrypted traffic has hidden payload information, side channel information like packet length, direction, and order of traffic between Amazon Echo and Cloud server are accessible [17]. Every voice command and its response have a unique encrypted traffic pattern that can be leveraged using Deep Learning algorithms. Voice Command Fingerprinting assumes the features packet length, direction, and order of each encrypted voice command are unique. As the content of the encrypted traffic is correlated with the voice commands and an attacker can use outgoing traffic (encrypted voice commands packets) and incoming traffic (encrypted response packets) to infer the user's voice commands [31].

Voice Command Fingerprinting may have unauthorized privacy disclosure when an attacker gets access to a user's voice command by analyzing encrypted network traffic. For example, an attacker who is able to identify the personal interests of a user can target advertisements or promotions to the user, which is a breach of user privacy. For example, if a user makes frequent commands related to sports, the attacker can promote sports gear or sports streaming subscription plans to the user. The consequences can be more severe. An attacker can leverage Voice Command Fingerprinting to determine the most frequent commands of the user and extend to other malicious attacks (such as Skill Squatting Attacks). Skill squatting can occur when a malicious skill has a similar name to that of another harmless skill, and the malicious skill gets triggered when the user commands Amazon Echo. The malicious skill can then be used to record user conversations, eavesdrop on sensitive information, steal passwords and credit card information, etc.

2.4 Deep Learning

Deep learning is a sub-class of machine learning where multiple layers are used to extract complex features from input data. Deep learning is widely popular for tasks like image processing, where lower layers may identify features like edges and colors while higher layers may identify overall patterns such as digits, human faces, letters, etc. Deep learning is primarily based on artificial neural networks, which are made up of a collection of artificial neurons. Artificial neurons loosely model the pattern of biological human neurons; a neuron can process and transmit signals to another neuron.

2.5 Ensemble Learning

Ensemble learning is a technique used in machine learning problems that involves combining multiple machine learning models to increase the effectiveness of a specific computational problem. The primary objective of ensemble learning is to enhance the performance of a model and minimize the risk of selecting a subpar model for tasks such as classification, prediction, or function approximation. Additionally, ensemble learning can be applied to assign confidence to model decisions, select optimal features, perform data fusion, facilitate incremental learning, adapt to non-stationary environments, etc.

CHAPTER III

LITERATURE REVIEW

The security and privacy of Amazon Echo is one of the most crucial aspects of the device. While users enjoy the features provided by the device, many often overlook the security and privacy implications of the device. They are unaware of the underlying security and privacy mechanisms due to which they may engage in activities that compromise their identity or data. Moreover, a device vulnerability may create an opportunity for an adversary to carry out an attack or manipulate users. Thus, such vulnerabilities are an area of interest to study and possibly offer mitigation. Each vulnerability can be classified into one of three categories, i.e., Software, Hardware, and System, according to its nature for a systematic study. Some portions of this section are reused from [25], the published work of the author itself.

3.1 Software Vulnerabilities

3.1.1 Skill squatting attack

Skills are the voice-driven capabilities developed for Alexa to power Amazon devices, such as Echo [1] that enrich the device's capabilities. A common skill usage scenario is illustrated in Figure 3.1. Though skills extend Alexa functionality, they introduce new attack vectors.

Skill squatting attack exploits predictable errors, including homophones, compound words, and phonetic confusion, to wrongly direct users to malicious skills. Attackers create malicious skills with a similar invocation and intent name to legitimate skills [23]. A user intending to access a benign skill may be routed to a malicious skill due to phonetic confusion. When a malicious skill gets access to the user's device, further attacks can be carried out from there. Skill squatting attack is comparable to domain name typo-squatting in web applications where domain name's common

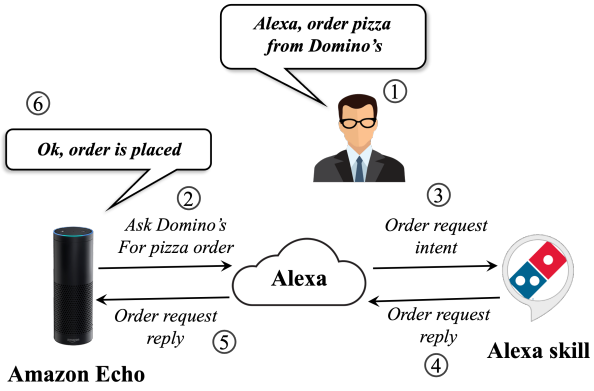

Figure 3.1: A User-Alexa interaction to order a pizza [19]

typos are exploited. Skill squatting vulnerability can be mitigated by adding a screening process during skill certification to scrutinize whether a skill can be confused with another registered skill. Currently, there are 30 skills with the name "Cat Facts", however, the mechanism of how Amazon routes a request is unknown.

3.1.1.1 Mitigation. Such vulnerable skills can be mitigated by thorough scrutiny at the screening process such that each skill has a unique name. Though this mechanism may mitigate the vulnerability, skill publishers may have a conflict over skill names. They may want a simple skill name which may lead to name scarcity.

3.1.2 Voice Masquerading Attack

In Voice Masquerading Attack (VMA), users are unaware of skill eavesdropping on their conversations. As a result, an adversary can exploit the vulnerability to extract a user's private information. There are two major types of VMAs [34], namely, In-communication skill switch and Faking termination.

In-communication skill switch is an opportunistic attack where a skill pretends to be another skill. The attack may occur when a user tries to switch skills during interaction with Alexa. A malicious skill pretends to hand over execution to the target skill by impersonating the target skill. As a result, the user may share the information intended for the target skill with malicious skill, which causes a serious privacy concern. Additionally, an adversary can exploit the acquired personal information to attack the user in the future.

Table 3.1: Summary of Amazon Echo (Software, Hardware, and System) Vulnerabilities [25]

Exploitation Mechanism	Vulnerability	Threat	Mitigation
Software	Skill Squatting Attack [23, 19]	Malicious skill gets control of device	Screening of new skill's name using Word-based and phenom-based techniques
	Voice Masquerading Attack [34]	Malicious skill eavesdrops user's communication	Skill response checker and User intention classifier
	Network Traffic Analysis Vulnerability [2]	Adversary can detect user-device interaction time	-
	BlueBorne attack vector [3]	Linux kernel and SDP server threats	Amazon published security patches
	Broadcast Media Vulnerability [22]	Echo triggered by broadcasting events	On-the-cloud system to detect media audios
	Automatic Speech Recognition Errors [9]	Alexa misunderstands words and triggers	Command discarded after looking on Amazon server
	Lack of Authorization Mechanism [21]	Any person can command Alexa	User-voice authentication mechanism
	Cross-Site Scripting Vulnerability [4]	Access, install and remove user's skills list	Findings shared with Amazon and issue fixed
Hardware	Dolphin Attack [33]	Inject inaudible commands using ultrasonic channel	Utilizing non-linearity traces that can not be erased during signal modulation
	Booting into Device Firmware [5] [12]	Echo can be exploited by gaining root shell access	Issue fixed in later iterations of Amazon Echo
System	Always listening Mechanism [13]	Alexa records and streams conversation without utilizing wake word	Turing Echo mic off while not using the device
	Lack of Physical Presence Detection Mechanism [21]	Echo picks up commands from outside window/door	VSButton to check physical presence of user

Table 3.2: Survey responses of Amazon Echo users [34]

Indicator of end of conversation	Users
Echo says "Goodbye" or something similar	23%
Echo does not talk	52%
The LED light on Echo is off	25%

Faking termination is a VMA where malicious skill fake skill termination to eavesdrop on a user. Users may rely on the skill's response to determine skill termination. For instance, users infer skill termination if the skill prompts "goodbye" or remains silent after execution. A list of users' perceived indicators of the end of a conversation is summarized in Table 3.2. Malicious skills may create fake termination while keeping eavesdropping on sensitive information of Amazon Echo users.

3.1.3 Broadcast Media Vulnerability

In January 2017, a six-year-old girl from Dallas accidentally ordered a dollhouse while playing with Amazon Echo [22]. The device ordered a dollhouse when the girl asked Echo, "Can you play dollhouse with me and get me a dollhouse?". Later, Echo devices in multiple households were triggered when a morning show covered the event. The Amazon Echos listening to the news tried to order a dollhouse.

3.1.3.1 Mitigation. In response to such events, Amazon developed an on-the-cloud system to distinguish media audio. The system uses broadcast audio to teach Alexa about recorded instances of Alexa's trigger words and use this knowledge to detect recorded sounds in the future. In addition, the system utilizes a technique called acoustic fingerprinting, an efficient mechanism that is robust to audio distortion and interference produced by television and other digital devices [26]. However, some false positive observations were detected when fingerprint match was tested on several videos [28]. In addition, some videos without a wake word had a fingerprint match, raising concern about the robustness of the technique.

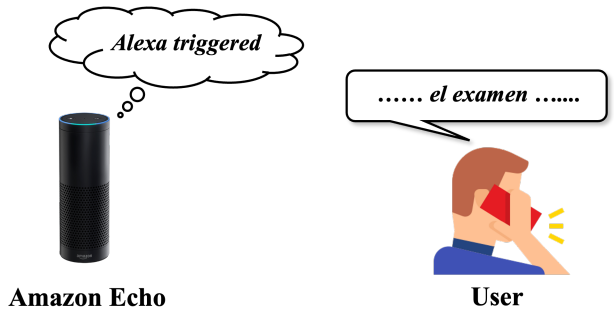

Figure 3.2: Speech recognition error: “El examen” interpreted as “Alexa” triggers the device

3.1.4 Automatic Speech Recognition (ASR) Errors

Though Amazon tried hard to address Alexa’s broadcast media vulnerability, Alexa is still vulnerable to various automatic speech recognition errors. Castell-Uroz *et al.* [9] experimented with an audio database with a few interesting findings.

A Spanish language audio database (approx. 1700 files) was employed to surveil Echo’s reaction to distinct sounds. Database audio was reproduced nearby Echo, where some words (e.g., “el examen”, “economia”) from the database triggered Alexa due to speech recognition error. A situation where Alexa got triggered due to an ASR error is shown in Figure 3.4. After getting triggered, Alexa looked up commands on the Amazon server but was eventually discarded. The results designate that Amazon’s security mechanism discards this kind of false positive.

3.1.4.1 Mitigation. Due to the limitations of speech recognition technology, ASR errors are unavoidable. However, attempts are made to minimize errors and improve performance. For example, Swarup *et al.* [30] diminished ASR errors by enhancing existing baseline model architecture with learned features. Similarly, Wang *et al.* [32] injected noise into error-free ASR-generated text data to train the dialog model with augmented data. The authors claimed to make VPA robust to ASR errors.

3.1.5 Network Traffic Analysis Vulnerability

There are multiple works on network traffic analysis of Amazon Echo in the literature. Apthorpe *et al.* [2] carried out a study in that direction by setting up a laboratory smart home

environment. In the experiment, Amazon Echo was asked a series of questions to observe the device's network traffic. The authors were able to identify the instances of user-device interactions using network traffic data. The knowledge of the user-device interaction time with an adversary may have unwanted implications and privacy concerns.

IVA and IVA-enabled devices mostly communicate over a secure channel using encrypted HTTPS [11]. However, such encryption cannot protect specific communication patterns like payload sizes, data rates, and source/destination. Many state-of-art machine learning techniques can leverage such information to infer user behaviors such as duration of user-device interaction, listening to music, and ordering products or services. In addition to that, machine learning algorithms may be used to predict user commands [17, 31].

3.1.6 Lack of Authorization Mechanism

A user commands Echo by speaking out a trigger word. The trigger word is "Alexa" by default, however, it can be configured to be one of the "Amazon", "Computer", or "Echo" [7]. There is an absence of an additional authentication layer to control access to the device, which is a serious vulnerability. Amazon Echo does not check if a command is issued by an authorized user or someone else, making it vulnerable to attackers who manage to get access to the device. In addition to that, Amazon Echo can be triggered by machine-generated voices due to the lack of an authentication mechanism [21]. MP3 audio files generated via an online resource have successfully accessed the device and executed commands. MP3 audio from various devices, such as Bluetooth speakers, laptops, desktops, and mobile phones, is capable of issuing commands to Alexa.

3.1.6.1 Mitigation. A layer of authentication can be implemented by adapting a biometrics-based authentication scheme in Amazon Echo. A camera module can be integrated to identify users and help enforce authentication schemes. Authorized users are verified by a face-recognition system when they gaze into the device [29]. A face-recognition algorithm wakes up the camera and authenticates users enabling a secure authorization mechanism. Once a user is authenticated, echo can listen and execute user commands securely. The biometric-based authentication can be implemented in future models of Amazon Echo. However, it is challenging to implement the authenti-

cation procedure in the current and previous models in the user households due to the hardware nature of mitigation.

3.1.7 Bluetooth Associated Vulnerability

IoT devices, including Amazon Echo, can be vulnerable to Bluetooth-associated vulnerability, which may compromise the device and user data. Additionally, Bluetooth-enabled devices are vulnerable to a "BlueBorne" attack vector that endangers the integrity of digital devices [3]. BlueBorne attack vector has eight zero-day vulnerabilities critical to IoT device security. Specifically, there are two vulnerabilities of Amazon Echo:

- Linux kernel: Remote code execution vulnerability
- SDP server: Information leak vulnerability

BlueBorne permits attackers to compromise a device even when Bluetooth is not in discoverable mode. For Amazon Echo, there is an absence of a mechanism to turn Bluetooth off given the device's limited user interface, making it vulnerable to BlueBorne attack. Additionally, Echo devices constantly scan for Bluetooth communications, increasing the attack risk.

3.1.7.1 Mitigation. Armis Labs apprised Amazon regarding BlueBorne attack vector-associated risks. Amazon issued an update in response to security fixes. In addition, Amazon Echo users (version>v591448720) have been automatically updated with the security patch.

3.1.8 Cross-Site Scripting Vulnerability

Cross-Site Scripting is an injection attack where malicious scripts are injected into harmless websites. Alexa can be vulnerable to Cross-Site Scripting (XSS), according to a study in August 2020 [4]. A Cross-Origin Resource Sharing (CORS) token can be extracted using XSS that is exploited to perform actions using the victim's identity. The attack shown in Figure 3.3 is carried out as follows:

- The user receives a malicious link with code-injection capability that redirects the user to Amazon. The user clicks on the malicious link.
- An AJAX request using the user's cookies are sent to access the list of the user's installed

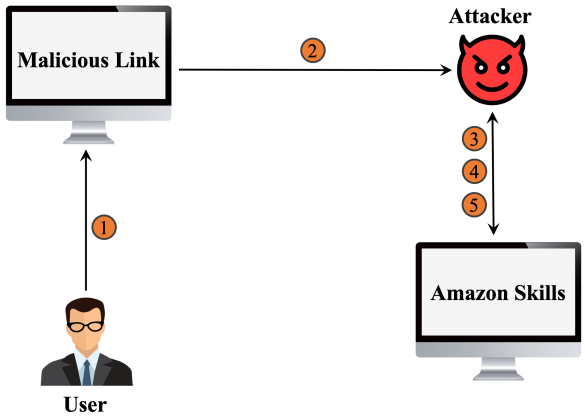

Figure 3.3: Attack flow using XSS and CSRF token

skills on his/her Alexa account. The CSRF (Cross-Site Request Forgery) token is retrieved as a part of the response.

- CSRF token is misused to remove a skill from the user's list of installed skills.
- Attacker now installs a skill whose invocation phrase is identical to the deleted skill.
- Malicious skill is triggered when the user uses an invocation phrase.

An adversary can exploit certain vulnerabilities in Alexa sub-domains to carry out attacks targeting Alexa users. Adversary takes advantage of these vulnerabilities to carry out multiple actions in multiple stages to attack targeted users. The attack initiates when the user clicks on a malicious link. An attacker can carry out the following attacks [24]:

- Access user's Alexa voice history.
- Install skills to the user's Alexa without the user's knowledge.
- View the list of users' Alexa skills.
- Remove a user's skill without the user's knowledge.
- Access user's personal information that includes bank details, personal details, addresses, phone numbers, etc.

3.1.8.1 Mitigation. The findings of the study illustrating the vulnerabilities were shared with Amazon. Amazon responded to it by fixing issues and pushing updates. No manual update is required from Echo users to mitigate the vulnerability.

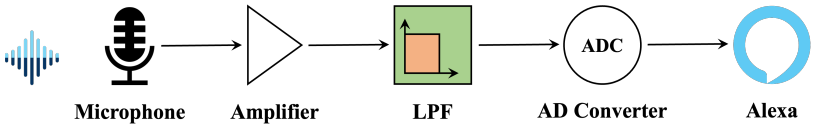

Figure 3.4: Demonstration of modulated tone passing through the signal pathway of an audio device in terms of FFT [33]

3.2 Hardware Vulnerabilities

3.2.1 Dolphin Attack

A Dolphin attack is an inaudible attack that exploits the ultrasound channel and underlying hardware vulnerability to inject inaudible voice commands at VPAs. The attack uses modulated audio commands on ultrasound carriers (frequency >20 kHz), making the command inaudible to the human ear [33]. The modulated command is demodulated and interpreted at voice capture hardware and speech recognition system, respectively, at VPA. The modulated audio signal can be successfully demodulated by leveraging the non-linearity of microphone circuits. Modulated signal traversing an audio capture device is illustrated in Figure **??**. The attack exploits Micro Electro Mechanical Systems (MEMS) microphones that accept inaudible ultrasound signals as legitimate commands. Since the attack employs synthesized ultrasound signals, an attacker requires proximity to the target device. For example, Amazon Echo can pick up and execute inaudible audio commands from a distance of 165 cm. The attack range was further increased to 25ft by exploiting the non-linearity of the Echo's microphone [27].

3.2.1.1 Mitigation. Dolphin attacks can be abused to carry out unsolicited actions on Amazon Echo. Therefore, defense strategies should be employed to address unwanted attacks. Hardware-based defense strategies such as microphone enhancement can be an approach in that direction. Since the current MEMS microphones can sense high frequency (>20 kHz) signals, they can be enhanced to suppress such signals. Similarly, there have been defense attempts utilizing the non-linearity traces, which cannot be erased during the signal modulation [27].

3.2.2 Booting into Device Firmware

Amazon Echo can be exploited physically, allowing an adversary to gain root shell access to the underlying Linux OS. Amazon Echo has two underlying vulnerabilities. [5]:

- Exposed debug pads at its base.
- Hardware configuration setting that permits booting the device via an external Secure Digital (SD) card.

These vulnerabilities can be exploited and allow the attacker to boot into the underlying Linux environment from an SD card [12]. Furthermore, an attacker can boot into the device's firmware and install a persistent backdoor that allows remote root shell access to the device. After the root access is obtained, the attacker can install malware, steal authentication tokens, and wiretap the device remotely. Rooting Amazon Echo requires physical access to the device, which may not be a concern for a device in a secure location such as a personal household. However, adapting Amazon Echo to places such as hotel rooms provides an avenue for attacking [16].

3.3 System Vulnerabilities

3.3.1 Always Listening Mechanism

Studies have shown that Amazon Echo starts recording and transmitting audio only after it gets triggered with a wake word [14]. Till then, it stays in a dormant state of buffering and re-recording until a wake word is detected. Ford and Palmer [13] carried out an experiment in that direction where they analyzed Echo Dots' network traffic over 21 days in a private household. Nobody in the household interacted with the devices on purpose, utilizing a wake word during this period. Analyzing the logged audio reveals that 70% of logged response cards were Television sounds and 30% were human voices. This demonstrates that Amazon Echo records private conversations without utilizing a wake word. This can be a significant privacy concern where personal or sensitive audio is leaked accidentally or by an attacker.

3.3.1.1 Mitigation. The vulnerability can be mitigated by turning the device mic off with a physical mechanism while speaking out private information. Alexa does not stream audio to

Figure 3.5: How to mute Amazon Echo? Echo's LED light turns red while the mic is off.

Amazon AVS cloud while the device mic is turned off [13]. The echo light turns red when the microphone is turned off, as shown in Figure 3.5. However, many users do not use the mic button despite being aware of the functionality. Multiple users perceive the technique negates the device's hands-free accessibility [20].

3.3.2 Lack of Physical Presence Detection Mechanism

Amazon Echo does not require a user to be physically present near the device to request a service. Due to the absence of a mechanism to detect the physical presence, an Alexa-enabled device executes any command that it can hear, provided that the command is loud enough. Any service request that reaches Amazon Echo at 60dB (or higher) sound pressure level gets served by the device. It is a severe vulnerability that can be exploited in multiple ways. For instance, an adversary can issue a command from the facade to access Amazon Echo inside a household. The adversary can then aggravate the attack by utilizing other devices connected to the Echo. Alternatively, an adversary can control Echo if he gets access to one of the speakers in proximity to the Echo in the household. The attacker can abuse the speaker to play audio containing wake words and commands to compromise Alexa-enabled devices.

3.3.2.1 Mitigation. A user's physical presence can be detected by the Virtual Security Button (VSButton), which is an access control technique that utilizes the physical presence of a

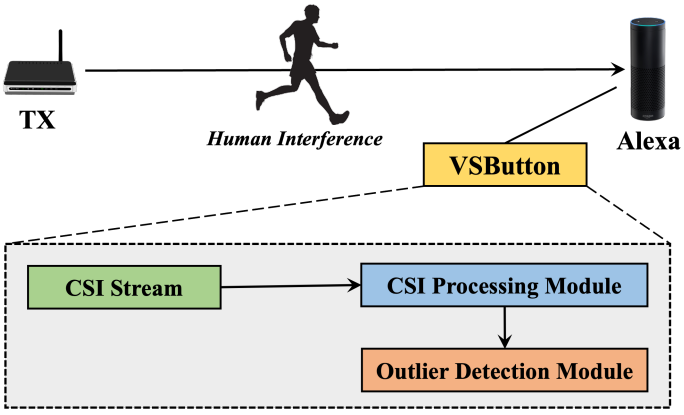

Figure 3.6: Design of VSButton [21].

user. The secure access mechanism allows access to Alexa only when VSButton is in a push state. The virtual button is pushed whenever a human presence is detected nearby. The access control mechanism utilizes a home WiFi network to detect user movement. VSButton monitors the Channel State Information (CSI) of home WiFi to detect human motion. A user can push VSButton simply by waving his hand. The variation in CSI values within a room can be leveraged to detect human motion. Movements inside a room cause considerable variation in CSI values, while movement outside the room/house causes only a tiny variation. The phenomenon is employed to determine if movement is occurring inside the room. The human movement detection by VSButton consists of two major steps:

- CSI processing phase;
- Outlier detection phase.

In CSI Processing Phase, noises in CSI values are eliminated. The output is then utilized in Outlier Detection Phase to detect CSI patterns of movements inside the room. A real-time hyper-ellipsoidal outlier detection mechanism is employed in the later phase to detect human movement. The components of VSButton are shown in Figure 3.6.

CHAPTER IV

VOICE COMMAND FINGERPRINTING ATTACK

4.1 Threat Model

The threat model of the voice command fingerprinting attack is shown in Figure 4.1. We assume that there is an adversary who can sniff the smart speaker's encrypted network traffic. For example, an adversary can be anyone who can eavesdrop victim's WiFi network and access network traffic between a smart speaker and a home Internet service provider server. The adversary is passive and can not modify packets in any way. In addition to that, the adversary can not decrypt encrypted packets. A possible real-world adversary can be Internet service providers or local network eavesdroppers.

We also consider that the smart speaker is Amazon Echo 2nd generation which is the highest selling smart speaker in the market [8]. The adversary has information on the model of the smart speaker. He can deduce the IP address of both the Amazon Echo and the Amazon servers that run the voice services. Note that an adversary can separate the network traffic of other devices connected to the same WiFi if he has the IP address of the Amazon Echo.

We make the assumption that packets sent to the server are considered outgoing packets, typ-

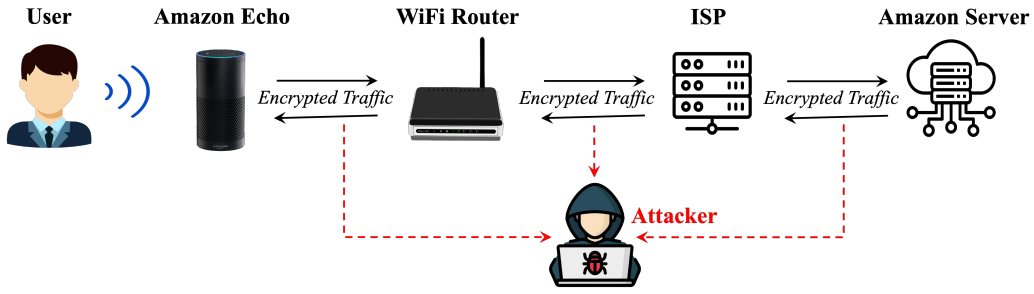

Figure 4.1: Threat mode of voice command fingerprinting [31].

ically containing voice commands, while packets sent to a smart speaker are considered incoming packets, often containing responses. We also assume that an attacker possesses the ability to deduce the start and end times of each traffic trace. In this scenario, the attacker can gather side-channel information such as direction (outgoing or incoming), packet size, and timestamp.

4.2 Neural Networks

Neural networks are inspired by the structure and mechanism of the human brain, where neurons take single or multiple inputs and process them before passing the output to the next layer of neurons. Neural networks contain layers of nodes that may be an input layer, hidden layers, and an output layer.

Each node also referred to as an artificial neuron, is interconnected with others and possesses a weight and threshold. When the output of a node exceeds the threshold value, it becomes activated and transmits data to the subsequent network layer. Else no data is forwarded to the next layer if the threshold is not met. In this manner, data moves through the network as each neuron processes and transmits information to its subsequent layer. Neural networks are considered the heart of deep learning algorithms. Neural networks can offer a diverse set of effective techniques in domains, including pattern recognition, data analysis, etc.

4.2.1 Convolutional Neural Networks (CNN)

A convolutional neural network, or CNN, is a type of artificial neural network. CNN is used popularly in recommendation systems, image/video recognition and classification, natural language processing, etc. CNNs have been previously utilized in website fingerprinting attacks with significant success, which indicates that they might be useful in voice fingerprinting attacks as well. CNN primarily consists of three types of layers [15]:

- Convolutional layer
- Pooling layer
- Fully-connected layer

In a CNN, a convolutional layer is the first layer, which may have additional convolutional

layers or pooling layers as the following layers. Generally, fully-connected layers are the final layers in a CNN. As more layers are added to CNN, the complexity increases, enabling it to identify greater portions of images. Earlier layers work on uncomplicated features such as color and edges. As the image is processed through CNN, larger portions and shapes are recognized.

4.2.1.1 Convolutional Layer. The Convolutional layer is considered the core building block of CNN that does the majority of computation operations. To carry out a computation, it needs a few components, such as input data, a filter, and a feature map. There is also a feature detector which is a kernel or a filter that can carry out the convolution procedure.

Feature detector, which is part of a two-dimensional array, represents part of an image. The filter size is variable but generally is a 3x3 matrix that determines the receptive field's size. The image is subjected to the filter, and the dot product is computed between the input pixels and the filter. This computed dot product is used as input for an output array. Subsequently, the filter moves by a certain stride, and the process is repeated until the filter has covered the entire image. The final result obtained from the sequence of dot products between the input and the filter is called a feature map or activation map. CNN applies a ReLU (Rectified Linear Unit) transformation after each convolution operation.

4.2.1.2 Pooling Layer. Pooling layers are the downsampling layers responsible for dimensionality reduction, which is done by reducing the input's number of parameters. Though the pooling operation applies a filter across the entire input (as in the convolutional layer), the filter does not have weights (contrary to the convolutional layers). Rather than that, the kernel performs an aggregation function on the values within the receptive field and fills the output array with the result. Primarily, there are two pooling categories:

- Max Pooling: Max pooling is the more popular pooling technique among the two categories. It chooses the pixel with the highest value and transfers it to the output array.
- Average Pooling: While traversing the input, the filter computes the mean value within the receptive field and transmits it to the output array.

4.2.1.3 Fully-Connected Layer. There is no direct connection between a pixel value of

the input image and output layers in partially connected layers. However, every node in the output layer of a fully connected layer is linked to a node in the preceding layer. A fully connected layer performs classification tasks based on extracted features from previous layers and filters. A fully connected layer usually uses a softmax activation function for classification, contrary to the ReLU function used by convolutional and pooling layers.

4.2.2 Long Short Term Memory (LSTM)

Long Short Term Memory is a category of recurrent neural network (RNN). A recurrent neural network is an artificial neural network built specifically to handle sequential data by preserving internal memory or state. Unlike feedforward neural networks that process input data in a linear manner, RNNs incorporate feedback connections that enable information to flow from one-time step or unit to the next within a sequence. Though RNN is highly effective, it has a shortcoming in handling "long-term dependencies". That's where the LSTM comes into play.

LSTM is a particular type of RNN capable of learning long-term dependencies. Initially introduced by Hochreiter & Schmidhuber in 1997, LSTM was refined and popularized by many succeding scientists. Long Short-Term Memory (LSTM) Networks address a limitation of standard RNNs, namely the issue of vanishing gradients. This problem arises when backward propagation of weights through the network causes the partial derivative of the loss function (gradients) to diminish (approach zero) as it progresses towards layers closer to the input. Consequently, the model's ability to learn from these initial layers is reduced. In the context of temporal sequences, this phenomenon results in the neglect of earlier parts of the sequence, commonly referred to as Short-Term Memory. Within the domain of smart speakers, this specifically relates to the segment within the trace where the user provides a voice command to the device.

To address the challenge of Short-Term Memory, Long Short-Term Memory (LSTM) networks employ two mechanisms: gates and cells. Cells act as memory units that store important information, such as significant signals from earlier points in a sequence. Gates control the storage of data in the cells, determine which data should be discarded, and regulate the flow of data entering or leaving the cells, connecting them with other units in the network. In the designed LSTM

network for this attack, we utilize sequences of LSTM layers followed by dropouts. The classification component of the model consists of a densely connected layer with a total number of units corresponding to the number of classes. The softmax activation function is applied to this layer.

4.3 Ensemble Learning

Ensemble learning uses multiple learning algorithms to achieve a better predictive capacity than obtained by any of the constituent individual algorithms alone. The principle is to combine the predictive capacity of each constituent model, also called weak learner, to form a single optimal strong learner. The weak learners are trained on the training set to generate individual predictions, and the final prediction outcome is determined by aggregating the results from all the weak learners.

4.3.1 Weighted Average Ensemble

A weighted average ensemble is an ensembling approach that combines the output from multiple base models to make a final prediction. The contribution of each base model in weighted average ensemble learning is weighted proportionally to the particular model's capability. That means the base model with higher predictive power is more important and assigned a greater weight for making a final prediction.

Given a set of N individual base models denoted as M1, M2, ..., MN, and their corresponding predictions for a given instance x represented as P1(x), P2(x), ..., PN(x), the weighted average ensemble prediction Y(x) can be calculated as:

$$Y(x) = w1 * P1(x) + w2 * P2(x) + ... + wN * PN(x)$$

Here, w1, w2, ..., wN represents the weights assigned to each individual base model. These weights can either be predetermined or learned from the training data. They determine the amount of contribution each model has on the final prediction, with larger weights indicating a greater impact. The sum of weights above must fulfill certain criteria, such as summing up to 1 (i.e., w1 + w2 + ... + wN = 1) to maintain proper normalization.

The weighted average ensemble approach enables diverse models with different capabilities

to contribute to making the final prediction. Thus, the final prediction is more comprehensive and accurate compared to that of a single model. A weighted average ensemble is particularly effective when the individual base models are diverse and possess complementary expertise. The ensemble model can leverage their unique strengths by combining the predictions of these models, resulting in improved performance.

Additionally, a weighted average ensemble is suitable for noise reduction, where models with varying levels of noise can be weighted appropriately to generate smoother and more robust predictions. It is also efficacious in decision fusion scenarios, where the objective is to integrate information from multiple sources in order to make well-informed decisions. Furthermore, it can be advantageous in situations where selecting the single best model is challenging, as it mitigates the risk of relying solely on a single model by combining results from multiple models. In summary, careful weight assignment, model diversity, and individual model quality are imperative factors in attaining optimal results when employing a weighted average ensemble.

4.3.2 Stacking Ensemble Learning

Stacking ensemble learning (also called stacked generalization) involves training a model (final estimator) to combine the predictions of multiple other models(base estimators). Firstly, base estimators are trained on the available data, then a combiner final estimator algorithm is trained on the cross-validated predictions of base estimators to make a final prediction that can prevent overfitting. An arbitrary combiner algorithm is used for stacking, although a logistic regression algorithm is used generally.

Figure 4.2 shows how the prediction of three base classifiers gets stacked to train the meta-classifier, which makes the final prediction. Here, three classifiers (*C1, C2, and C3*) are individually trained and are used in the stack. The predictions (*P1, P2, and P3*) are then used as input for the meta-classifier, which makes the final prediction. In stacked generalization, cross-validation is used to prevent overfitting while training base models, whose predictions are used to train meta-classifier. Additionally, stacking requires careful model selection, training, and validation to produce optimal results.

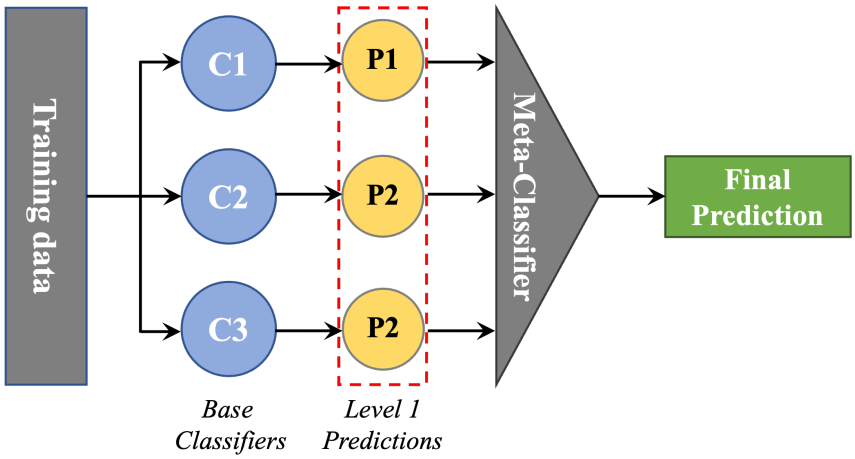

Figure 4.2: Schematic of stacking/stacked generalization[10]

CHAPTER V

EVALUATION

Three individual deep-learning models were utilized for carrying out the fingerprinting attack. In addition to that, different ensemble techniques were investigated to assess the performance and increase the effectiveness of the attack. The implementation details are presented in this chapter.

5.1 Dataset

The dataset collected by Wang et al. [31] was used for carrying out the attack. The dataset comprises 150,000 encrypted traffic traces from Amazon Echo 2nd generation devices, gathered between March 2019 and August 2019. The dataset includes traces of 100 voice commands; each voice command has 1,500 traffic traces. The voice commands were compiled from Amazon Echo weekly emails from December 2018 to March 2019. The emails contained popular voice commands issued by Echo users. The voice commands can be categorized into one of three categories:

- Single response command
- Time-sensitive command
- Multiple response command

Single response command is a voice command with an identical response from a server every time the command is issued. For e.g., the response to the command "*How deep is the Indian ocean?*" was always identical. Time-sensitive response command is a voice command whose response may change over time. For e.g., the response to the command "*What is the price of bitcoin?*" was different depending upon time. Multiple response command is a voice command whose response is one of the finite possible responses. For example, when issuing the command

"*Tell me a barbecue joke*", the Amazon server randomly selected and returned one joke from a fixed set of five jokes during data collection.

5.2 Experimental Setting

The proof-of-concept attack was implemented in Python 3.9, along with Keras at the front end and Tensorflow at the back end for neural network implementation. The attack was performed on a Linux machine (Ubuntu 22.04) with an Intel i7 - 9800X CPU @ 3.80GHz, 16 GB memory, and a GPU (NVIDIA GeForce GTX 1080). The experiment ran on a Conda environment by leveraging the deep learning model's library support for the NVIDIA CuDNN library, which greatly reduced the training time compared to the CPU's. The data was partitioned as follows: Training (64%), Validation (16%), and Testing (20%).

5.3 Results

All the attacks were implemented in a closed-world setting. The attack performance of three base models is summarized in Table 5.1. Similarly, the attack performance of weighted average ensemble models is presented in Table 5.2, and the performance of stacked generalization is presented in Table 5.3. Note that, for ensemble models, different combinations of three base models are implemented. For all experiments, the attack was run 5 times, and the final accuracy is presented as the mean accuracy of all runs. The difference in accuracy values across different runs is presented as the variance of the attack.

The best-performing model is CNN, with 88.35% accuracy. The LSTM model performs with 86.26% accuracy while SAE performs much lower at 74.69%. The results obtained are slightly less than that of the reference project [31], which might have been caused by the hardware differences. The average training times across all runs are also presented in the table.

5.4 Performance Impact of Ensemble Learning

Using ensemble learning increased the attack performance as shown in Table 5.2 and Table 5.3. We tried different ensemble combinations of the base models to gauge the performance of each combination. The highest performance of the weighted average ensemble is achieved when

Table 5.1: Performance of deep learning attacks

Deep learning model	Accuracy	Variance	Training time(Minutes)
CNN	88.35%	$1.35*10^{-5}$	61.10
LSTM	86.26%	$4.14*10^{-7}$	214.07
SAE	74.69%	$4.59*10^{-6}$	22.7

Table 5.2: Performance of Weighted average ensemble attacks

Ensemble base models	Accuracy	Variance
CNN & LSTM	88.66%	$1.52*10^{-5}$
CNN & SAE	88.78%	$2.15*10^{-5}$
LSTM & SAE	84.60%	$5.53*10^{-6}$
CNN & LSTM & SAE	88.51%	$5.79*10^{-6}$

CNN and SAE are used as base models. Apart from that, we achieved only 88.51% accuracy while using all three base models in the ensemble model. For the weighted ensemble, normalized weights were calculated using accuracy on validation data, while the attack accuracy is reported on the test data. When a is the validation accuracy, a normalized weight W for the base model was computed by the equation:

$$W_i = \frac{a_i}{\sum_{k=1}^{n} a_k} * 1$$

Stacked generalization has a better performance among all attacks. The highest result is obtained at 90.54% accuracy when all three base models are used in the ensemble. Stacked generalization increases the attack performance because it combines information from all base models and then trains a Logistic Regression on the combined information. This approach of ensembling by training a new machine learning algorithm on the combined data allows for a more complex classification of traffic data using all the features extracted by individual base models, giving a superior attack performance.

Table 5.3: Performance of Stacked generalization attacks

Ensemble base models	**Accuracy**	**Variance**
CNN & LSTM	89.84%	$8.82*10^{-6}$
CNN & SAE	88.99%	$4.80*10^{-6}$
LSTM & SAE	84.79%	$2.72*10^{-5}$
CNN & LSTM & SAE	90.54%	$1.03*10^{-5}$

CHAPTER VI

CONCLUSION AND FUTURE WORKS

In this research, a network traffic analysis attack, namely voice command fingerprinting, was implemented on encrypted network traffic data of Amazon Alexa. A few ensemble learning techniques were implemented to increase the attack performance. Using complex ensemble techniques in fingerprinting attacks can improve the effectiveness of the attack.

6.1 Conclusion

In this research, a comprehensive analysis of smart speaker security vulnerabilities is presented. Firstly, we introduce different security vulnerabilities of Amazon Alexa along with corresponding mitigation techniques. Then, a voice command fingerprinting attack is implemented that leverages ensemble learning techniques to analyze encrypted network traffic. The attack identifies the voice commands issued by users to their Amazon Echo devices. By combining predictions from multiple deep learning models, the ensemble attack model achieves superior performance compared to individual models. Stacking a generalization ensemble with three base models (CNN, LSTM, and SAE) accurately predicted 90.54% of voice commands in a closed-world setting. Our findings disclose a notable threat to smart speaker users who unknowingly reveal confidential information. The security and privacy concern may impact millions of smart speaker users worldwide.

6.2 Future Works

6.2.1 Defense

Since the VC fingerprinting attack is implemented, the top priority is to develop a suitable defense mechanism. A defense mechanism has been proposed previously [31] implementing adaptive padding and differential privacy. But the mechanism introduces aggressive overheads and

delays, which largely compromise usability. A delay of more than a few seconds causes poor user experience, and longer delays may result in timeouts where no response is delivered to the user. A technique like adversarial machine learning can be efficacious in this scenario. Adversarial machine learning can be implemented to make minimal perturbations to the traffic data so that deep learning models incorrectly classify them. Such a defense mechanism may not introduce overheads and delays, which can be a suitable countermeasure to the attack.

6.2.2 Real World Evaluation

The current evaluation has executed the attack under the scope of limited commands, which is not the case in real-world scenarios. The evaluation does not indicate how the performance scales when the command list is increased to match the real-world scope of such devices closely. Additionally, an evaluation in the open-world scenario needs to be carried out. In an open-world scenario, a traffic trace is determined to be either present or not in a list of the attacker's monitored traffic traces. This experimental setup was investigated in the website fingerprinting domain and can be studied in the voice command fingerprinting domain.

6.2.3 Attack on Other Smart Speakers

We conducted the attack on Amazon Echo 2nd generation smart speaker. There are several brands of smart speakers that were not evaluated in the experiment. Smart speakers like Google Home from Google and Homepod from Apple can be investigated for their vulnerability to the VC Fingerprinting attack as future work.

REFERENCES

[1] A. Alhadlaq, J. Tang, M. Almaymoni, and A. Korolova, *Privacy in the Amazon Alexa skills ecosystem*, Star, 217 (1902).

[2] N. Apthorpe, D. Reisman, and N. Feamster, *A smart home is no castle: Privacy vulnerabilities of encrypted iot traffic*, arXiv preprint arXiv:1705.06805, (2017).

[3] A. Armis, *BlueBorne Cyber Threat Impacts Amazon Echo and Google Home*, Armis.

[4] D. Barda, R. Zaikin, and Y. Shriki, *Keeping the gate locked on your IoT devices: Vulnerabilities found on Amazon's Alexa*, Check Point Research.

[5] M. Barnes, *Alexa, are you listening?*, F-Secure Labs.

[6] D. Bohn, *Amazon says 100 million Alexa devices have been sold - what's next?*, The Verge.

[7] ——, *You can finally say 'Computer' to your Echo to command it*, The Verge.

[8] A. Businesswire, *Amazon Echo Has 23% Share of Smart Speakers in Use: Strategy Analytics*, Businesswire.

[9] I. Castell-Uroz, X. Marrugat-Plaza, J. Solé-Pareta, and P. Barlet-Ros, *A first look into Alexa's interaction security*, in Proceedings of the 15th International Conference on emerging Networking EXperiments and Technologies, 2019, pp. 4–6.

[10] F. Ceballos, *Stacking Classifiers for Higher Predictive Performance*, Towards Data Science.

[11] H. Chung, M. Iorga, J. Voas, and S. Lee, *Alexa, can I trust you?*, Computer, 50 (2017), pp. 100–104.

[12] I. Clinton, L. Cook, and S. Banik, *A survey of various methods for analyzing the amazon echo*, The Citadel, The Military College of South Carolina, (2016).

[13] M. Ford and W. Palmer, *Alexa, are you listening to me? An analysis of Alexa voice service network traffic*, Personal and ubiquitous computing, 23 (2019), pp. 67–79.

[14] S. Gray, *Always on: privacy implications of microphone-enabled devices*, in Future of privacy forum, 2016, pp. 1–10.

[15] A. IBM, *Convolutional Neural Networks*, IBM.

[16] C. Jackson and A. Orebaugh, *A study of security and privacy issues associated with the Amazon Echo*, International Journal of Internet of Things and Cyber-Assurance, 1 (2018), pp. 91–100.

[17] S. Kennedy, H. Li, C. Wang, H. Liu, B. Wang, and W. Sun, *I can hear your alexa: Voice command fingerprinting on smart home speakers*, in 2019 IEEE Conference on Communications and Network Security (CNS), IEEE, 2019, pp. 232–240.

[18] S. M. Kennedy, *Encrypted traffic analysis on smart speakers with deep learning*, PhD thesis, University of Cincinnati, 2019.

[19] D. Kumar, R. Paccagnella, P. Murley, E. Hennenfent, J. Mason, A. Bates, and M. Bailey, *Skill squatting attacks on Amazon Alexa*, in 27th USENIX security symposium (USENIX Security 18), 2018, pp. 33–47.

[20] J. Lau, B. Zimmerman, and F. Schaub, *Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers*, Proceedings of the ACM on Human-Computer Interaction, 2 (2018), pp. 1–31.

[21] X. Lei, G. Tu, A. X. Liu, C. Li, and T. Xie, *The insecurity of home digital voice assistants-vulnerabilities, attacks and countermeasures*, in 2018 IEEE Conference on Communications and Network Security (C‘NS), IEEE, 2018, pp. 1–9.

[22] A. Liptak, *Amazon’s Alexa started ordering people dollhouses after hearing its name on TV*, The Verge.

[23] Y. Lit, S. Kim, and E. Sy, *A Survey on Amazon Alexa Attack Surfaces*, in 2021 IEEE 18th Annual Consumer Communications & Networking Conference (CCNC), IEEE, 2021, pp. 1–7.

[24] R. Nash, *Amazon Alexa virtual assistant security bug fixed after cybersecurity firm discovered vulnerabilities*, 8NewsNow.

[25] S. Pathak, S. A. Islam, H. Jiang, L. Xu, and E. Tomai, *A survey on security analysis of amazon echo devices*, High-Confidence Computing, (2022), p. 100087.

[26] M. Rodehorst, *Why Alexa won’t wake up when she hears her name in Amazon’s Super Bowl ad*, Amazon Science.

[27] N. Roy, S. Shen, H. Hassanieh, and R. R. Choudhury, *Inaudible Voice Commands: The {Long-Range} Attack and Defense*, in 15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18), 2018, pp. 547–560.

[28] L. Schönherr, M. Golla, T. Eisenhofer, J. Wiele, D. Kolossa, and T. Holz, *Exploring accidental triggers of smart speakers*, Computer Speech & Language, 73 (2022), p. 101328.

[29] B. Sudharsan, P. Corcoran, and M. I. Ali, *Smart Speaker Design and Implementation with Biometric Authentication and Advanced Voice Interaction Capability.*, in AICS, 2019, pp. 305–316.

[30] P. Swarup, R. Maas, S. Garimella, S. H. Mallidi, and B. Hoffmeister, *Improving ASR Confidence Scores for Alexa Using Acoustic and Hypothesis Embeddings.*, in Interspeech, 2019, pp. 2175–2179.

[31] C. Wang, S. Kennedy, H. Li, K. Hudson, G. Atluri, X. Wei, W. Sun, and B. Wang, *Fingerprinting encrypted voice traffic on smart speakers with deep learning*, in Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks, 2020, pp. 254–265.

[32] L. Wang, M. Fazel-Zarandi, A. Tiwari, S. Matsoukas, and L. Polymenakos, *Data augmentation for training dialog models robust to speech recognition errors*, arXiv preprint arXiv:2006.05635, (2020).

[33] G. Zhang, C. Yan, X. Ji, T. Zhang, T. Zhang, and W. Xu, *Dolphinattack: Inaudible voice commands*, in Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 2017, pp. 103–117.

[34] N. Zhang, X. Mi, X. Feng, X. Wang, Y. Tian, and F. Qian, *Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems*, in 2019 IEEE Symposium on Security and Privacy (SP), IEEE, 2019, pp. 1381–1396.

APPENDIX

APPENDIX

Table A.1: List of voice commands

Index	**Voice Command**
1	Are you wearing green?
2	Announce Happy Valentines Day.
3	Do dogs dream?
4	Do you like cats or dogs?
5	Flip a coin.
6	Give me a dinosaur fact.
7	Give me a fun fact about sleep.
8	Good Morning.
9	Help.
10	How deep is the Indian Ocean?
11	How do you spell appreciate?
12	How far away is the moon?
13	How hot is the sun?
14	How many days are in September?
15	How many days in a year?
16	How many days until Christmas?
17	How many days until Thanksgiving?
18	How many fantasy points does LeBron James have?
19	How many ounces in a pound?

Table A.1, cont.

Index	Voice Command
20	How many seconds are in a year?
21	How many teaspoons are in a tablespoon?
22	How much does an elephant weigh?
23	How much is an ounce of gold?
24	How old are you?
25	How old is Henry Winkler?
26	How old is Serena Williams?
27	How tall is Steph Curry?
28	How tall is the Empire State Building?
29	How tall is The Rock?
30	Is a tomato a fruit or a vegetable?
31	Pick a number?
32	Surprise me.
33	Talk like a pirate.
34	Tell me a barbecue joke.
35	Tell me a coffee joke.
36	Tell me a fun fact.
37	Tell me a Halloween hack.
38	Tell me a joke.
39	Tell me a palindrome.
40	Tell me a Star Wars joke.
41	Tell me some good news.
42	Tell me something weird.
43	Translate good morning to Spanish.
44	What are some power shops nearby?

Table A.1, cont.

Index	Voice Command
45	What are the most popular books this week?
46	What are the standings in the English Premier League?
47	What are you thankful for?
48	What can you do?
49	What happened in the midterm elections?
50	What is brief mode?
51	What is gluten?
52	What is Homecoming about?
53	What is my sports update?
54	What is my traffic report?
55	What is on your mind?
56	What is Roblox?
57	What is the AFC North Standings?
58	What is the best comedy movie?
59	What is the capital of Spain?
60	What is the date tomorrow?
61	What is the fourth book in the Narnia series?
62	What is the history of Labor Day?
63	What is the longest word?
64	What is the number one song this week?
65	What is the price of bitcoin?
66	What is the scariest movie of all time?
67	What is the score of the Eagles game?
68	What is the score of the Red Sox game?
69	What is the time in Singapore?

Table A.1, cont.

Index	Voice Command
70	What is the weather for Sunday?
71	What is the weather?
72	What is trending?
73	What is your favorite flower?
74	What is your favorite game?
75	What is your favorite hobby?
76	What is your favorite sport?
77	What is your mission?
78	What is zero divided by zero?
79	What movies are playing?
80	What were yesterday's scores?
81	When does daylight saving time end?
82	When does Game of Thrones return?
83	When is Boxing Day?
84	When is Hanukkah?
85	When is the NBA all-star game?
86	When is the next full moon?
87	Where did Yoda live?
88	Where is Mount Rushmore?
89	Who do you love?
90	Who is in Mastodon?
91	Who is nominated for best actor?
92	Who is playing Monday Night Football?
93	Who is second in the NBA Western Conference?
94	Who is winning the World Series?

Table A.1, cont.

Index	Voice Command
95	Who is your favorite author?
96	Who is your favorite poet?
97	Who is your favorite superhero?
98	Who scored for the Golden Knights?
99	Why do leaves change color in the fall?
100	Will it rain tomorrow?

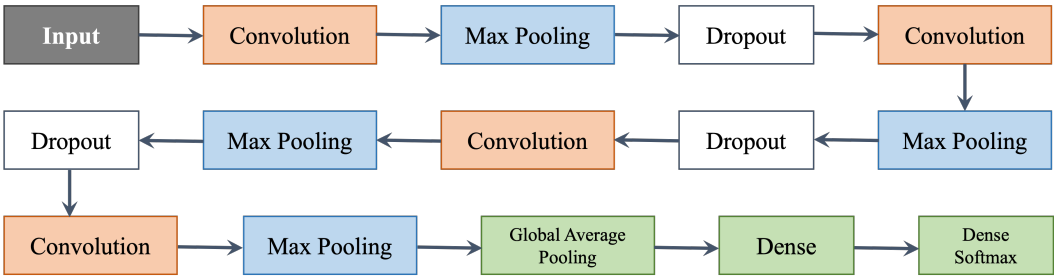

Figure A.1: Architecture of CNN Model

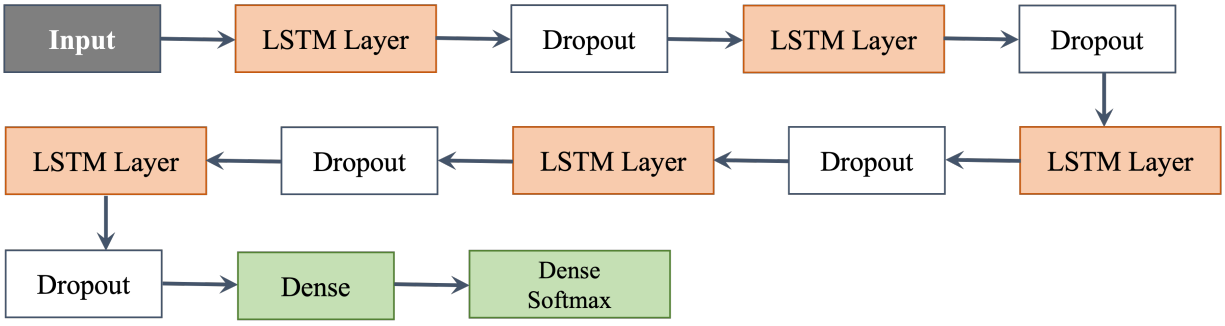

Figure A.2: Architecture of LSTM Model

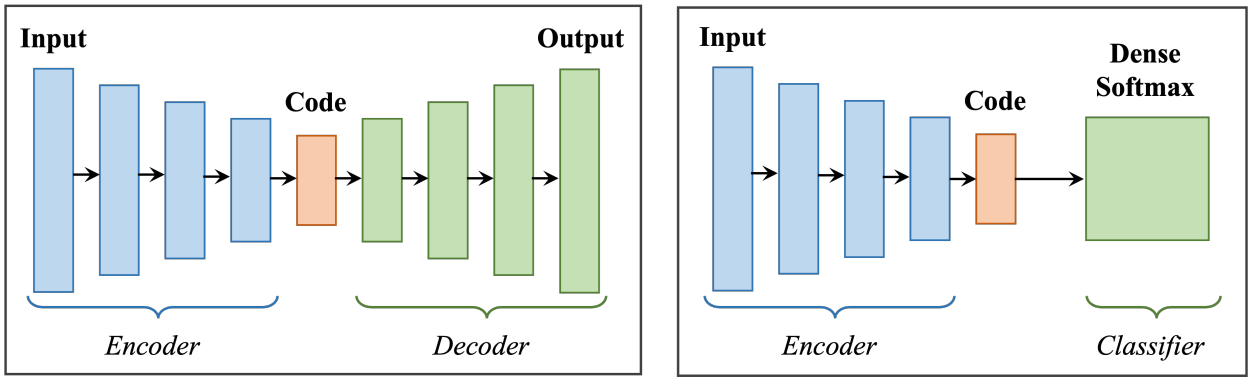

Figure A.3: Architecture of SAE Model

BIOGRAPHICAL SKETCH

Surendra Pathak was born in Nawalparasi, Nepal. He completed his Bachelor of Science in Computer Science and Information Technology from Tribhuvan University in 2017. He worked in the industry for a few years before resuming academic pursuits. Then, he enrolled in the Master of Science in Computer Science at The University of Texas Rio Grande Valley (UTRGV). He earned his Master of Science in Computer Science from UTRGV in July 2023. He can be reached at pathak.surendra01@gmail.com