Chapter 1: Introduction

Over the past two decades, the Internet has transcended human civilization to the information age. Cyberspace has now become a part of our ecosystem, an essential and indispensable element of our life. However, with the rise of Internet users, it has also become a playground of reprobaters who are constantly threatening the computer systems with ever evolving threats. These threats target various computing platforms like PCs, laptops, tablets, mobiles, IoT devices, etc. Security agencies globally are striving hard to combat these threats but fail to keep pace with their rising complexity and variety. Artificial Intelligence (AI), which heralded the growth of smart cognitive systems in the last decade, gives hope to cybersecurity agencies to contain cyber threats. This thesis is a step in this direction. The work done in this thesis is at the intersection of AI and Cybersecurity.

The following two sections, 1.1 and 1.2, describe the broad areas covered in the thesis, Cybersecurity, and AI, respectively. Subsequent sections describe the research gaps (1.3), the main contribution of the thesis (1.5), and the thesis organization (1.6).

1.1 Cyber Threats: The Need to Tame the Surge!

Cyber threats manifest themselves in various forms. They encompass attacks that seek to access information, damage data, deny information, or disrupt digital operations. They may be initiated by hacktivists, spies, terrorist groups, criminal gangs, nation-states, disgruntled employees, or lone hackers. Depending upon the origin and intent of the attack, cyber threats are classified as 'Cyber Crime', Cyber Terrorism', and 'Cyber Attack' [1]. Furthermore, based on the type of threat, they are typically categorized as shown in **Figure 1.1** [2].

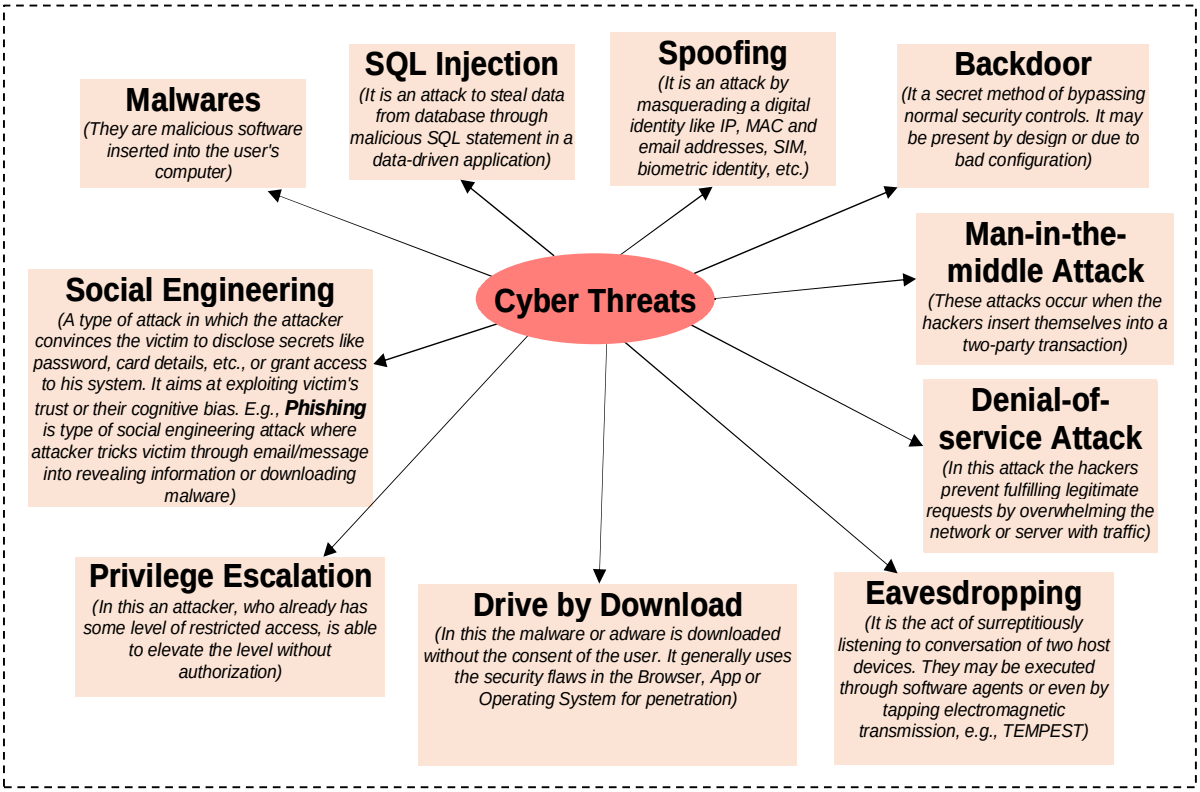

Figure 1.1: Types of Cyber Threats [2]

The cyber threats described in **Figure 1.1** above are rarely found in isolation. They are generally multi-vector and polymorphic, i.e., combine several types of attacks and change forms to avoid detection. In fact, in most episodes, no matter what kind of attack vector is used, it generally culminates with the insertion of malware during the exploitation stage. Thus, 'malware' is the most common and omnipresent cyber threat that affects most Internet users. Malwares can be of various types, as listed in Table 1.1. All these malwares can be inserted through means as already described in **Figure 1.1**. Amongst these means, 'Drive-by-Download' is a hacker's preferred medium of attack on web-based platforms [3].

Table 1.1: Types of Malwares

Virus:	*A self-replicating program that attaches itself to clean file and spreads throughout a computer system, infecting files with malicious code.*
Trojan:	*A type of malware that is disguised as legitimate software. Cybercriminals trick users into uploading Trojans onto their computer where they cause damage or collect data.*
Spyware:	*A program that secretly records what a user does, so that cybercriminals can make use of this information. E.g., spyware could capture credit card details.*
Ransomware:	*Malware which locks down a user s files and data, with the threat of erasing it unless a ransom is paid.*
Adware:	*Advertising software which can be used to spread malware.*
Botnets:	*Networks of malware infected computers which cybercriminals use to perform tasks online without the user's permission.*

Cybersecurity, which refers to protecting digital systems from cyber threats, can be categorized as shown in **Figure 1.2** [4].

Figure 1.2: Categories of Cyber Security [4]

Amongst these, 'Application Security' is the most critical due to large number of applications being used globally on different kinds of devices. Application security refers to security measured at the application level that

prevents data theft or injection of malicious code. Within the realms of application security lies web security and mobile app security. Both web security and mobile app security cover most of the aspects of application security, as most applications today are either web-based or mobile-based. The thesis is focused on web security and mobile app security, thus covering the most significant aspect of the cybersecurity paradigm, as it attempts to address the rising threat on web applications like browsers, hybrid mobile apps, etc.

1.1.1 Web Attack Vectors and Web Security

The number of users browsing the Internet using web-based applications has grown exponentially over the last few years [5]. Amongst the popular web applications are the browsers like Microsoft's Internet Explorer and Edge [6], Mozilla Firefox [7], Google Chrome [8], etc. Also, with the rising number of mobile users, mobile app users have increased. With the increasing popularity of mobile apps, 'Hybrid mobile apps', which use web-application-based protocols, are getting popular [9]. In fact, most popular apps on mobiles like Facebook, Twitter, Instagram etc., use the hybrid app technology [10]. With the rising number of these web platforms, the web is the preferred route for infecting connected devices. With billions of websites active on the Internet, hundreds of them added each minute, and most of them updating their content frequently, any web security expert seems to be walking in a labyrinth. Most of these web-based attacks are 'Drive-by-download' attacks [11], which inject malicious JavaScript from the server hosting the malicious web applications to the browser or the hybrid app. As per the latest Internet security reports, such drive-by-download attacks have increased manifold over the past years [12]. These attack vectors have also become smarter, making it difficult for older anti-malware technology to detect them. The older anti-malware techniques like static and dynamic heuristics [13] fail to detect most polymorphic and metamorphic malwares [14] (while static heuristics involve decompiling a program and testing its source code, in dynamic heuristics program is run in a controlled virtual environment to see the changes being made at runtime). Further, with automated tools being used to generate new ever-evolving malwares, this task is becoming more and more daunting [15]. The thesis attempts to overcome such limitations using AI for web malware detection.

1.1.2 Web Attack Vectors on Mobile Platform

Mobile is the most ubiquitous gadget on Earth today. The most common mobile platforms are Android and iOS, constituting 99.19% of the mobile ecosystem (72.72% Android and 26.47% iOS) [16]. The Android platform has emerged as the most popular computing platform that has more than three billion devices active across the globe [17]. These devices include not only mobiles and tablets, but even Android auto modules in cars, various Android versions running on televisions, watches, and a host of other smart and IoT devices. What makes things more challenging and interesting for the Android developers and security experts is that various versions of the Android Operating System, from Android 2.3.3 (Ginger Bread) to Android 11.0, coexist in this ecosystem [18]. The thesis has discussed threats that emanate from hybrid Android apps. These apps use WebViewcomponent for handling web content within Android apps [19]. WebView allows HTML and JavaScript to run and render webpages inside the apps, thereby allowing them to download content from web servers on the Internet. It is used by several popular apps, like Facebook, Twitter, Instagram, Gmail, Amazon, etc. [10].

1.2 Artificial Intelligence (AI)

Artificial Intelligence (AI) refers to the intelligence demonstrated by machines [20]. Basically, it describes cognitive behaviour like problem-solving through learning that mimics the human brain. The name traces its origin to the natural intelligence displayed by humans and animals, which, unlike AI, also comprises emotions and consciousness. The word was coined in academic circles way back in the year 1956 [21]. Since then, AI has been used differently in various sub-fields like robotics, medicine, earth sciences, etc., and machine learning and has gone through multiple cycles of disappointment and success. It rose to prominence in 2015 after AlphaGo defeated a professional Go player using Deep Mind's software based on deep learning [22]. This marked the re-emergence of Artificial Neural Networks (ANN) based Deep Learning (DL) technology and the whole AI paradigm. **Figure 1.3** gives the chronology of AI, Machine Learning (ML), and DL and their relationship [21]. It is noticeable that DL is a subfield of ML, which is, in turn, a subfield of AI.

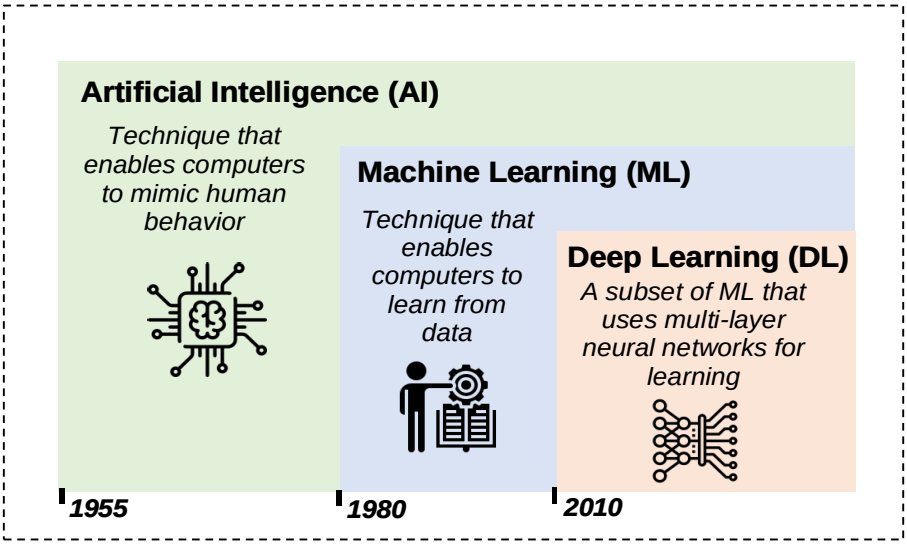

Figure 1.3: Evolution and Relationship of AI, ML, and DL

AI is divided into types based on two categories, viz., based on capability and functionality, as shown in **Figure 1.4** [23]. For example, AI currently available is narrow/weak AI (based on functionality categorization- reactive and limited memory machines).

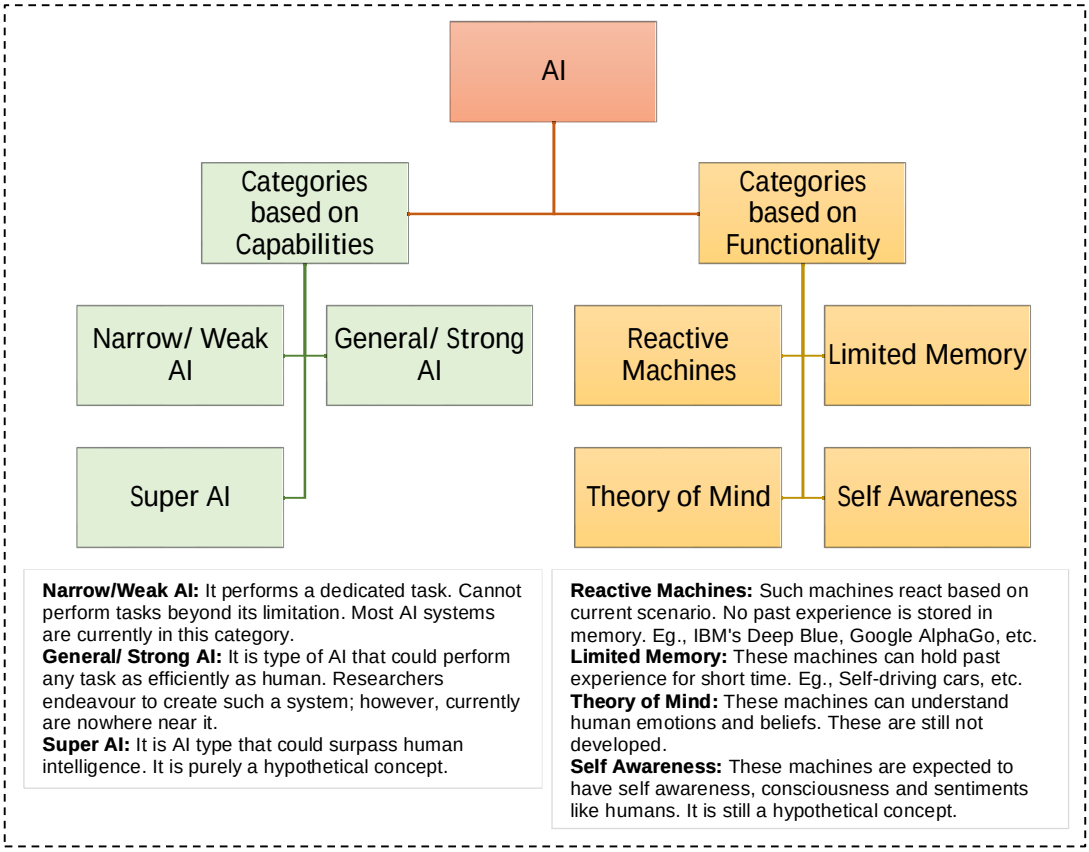

Figure 1.4: Categorization of AI

The narrow/weak AI, currently prevalent, has various components as illustrated in **Figure 1.5** [24].

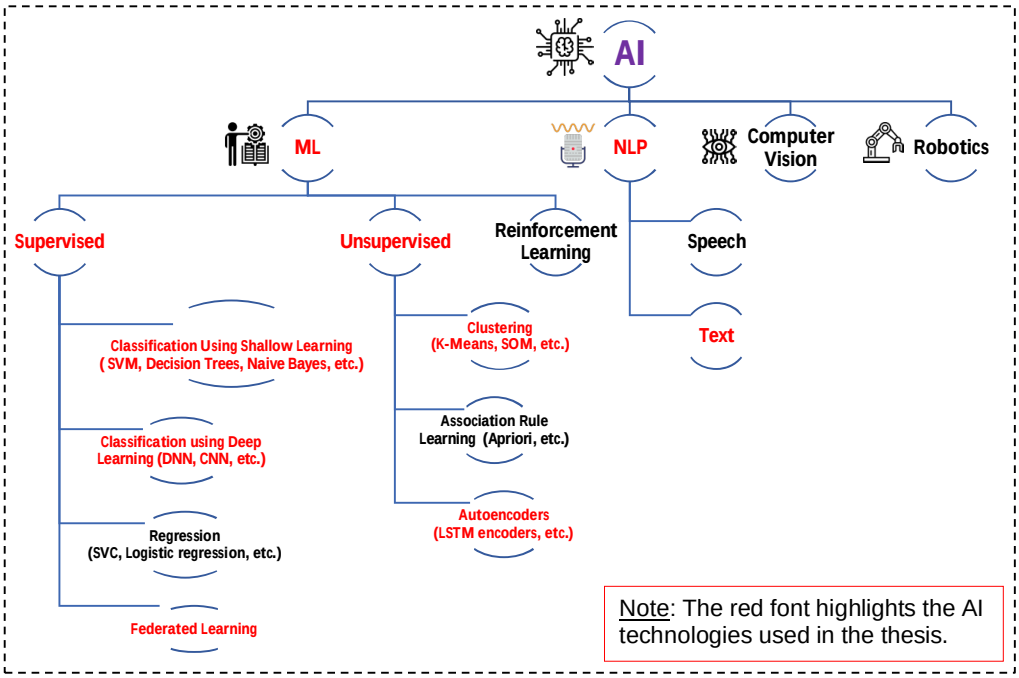

Figure 1.5: Components of Narrow/Weak AI [24]

AI is now being used in numerous fields, including self-driven cars, language translation, fraud detection, facial recognition, business recommendations engines, robotics, voice recognition, etc. The work done in this thesis has used AI in the field of web security. Various techniques of AI, as highlighted in red in **Figure 1.5**, have been utilized in the thesis to find solutions for web security related problems. The thesis attempts to utilize the complete spectrum of AI capabilities to solve some of the most pressing and critical problems related to web security.

1.3 Motivation & Research Gaps

Over the years, web security challenges have been increasing with evolving threats in both numbers and complexity. Concurrently, there have been promising and significant developments in the field of AI, and it is being applied to solve long-standing problems in science, engineering, and social sciences. The motivation for this thesis has been derived from these developments in web security and AI. We attempt to mitigate different kinds of web security threats by innovatively leveraging the latest techniques and

coming up with scalable and deployable solutions. We have identified research problems based on the following gaps/observations/challenges:

- *Challenges in Data Collection about Malicious Webpages:* To effectively train machine learning models, we need to collect a reasonable amount of data about malicious webpages. Malicious webpages, a major concern for web security, constitute a mere 0.1-1 % of the indexed web [25]. Further, the indexed web is a mere 4% [26] of the entire webspace on the Internet, including the notorious Darknet [27]. The collection of such malicious webpages and the malwares hosted on them is complex as these sites are difficult to find, and once found, they use evasive techniques to avoid detection by cybersecurity agencies. Unfortunately, no suitable mechanism exists for the collection of such webpages. We have addressed this gap by developing "MalCrawler", a customized and focused crawler that seeks more malicious webpages than a generic crawler and collects relevant data about webpages by overcoming the evasion techniques deployed by such sites (Chapter 2 has further details on this research gap).

- *Effective Classification of Malicious Webpages using Conventional and Deep Machine Learning*: Earlier techniques that were used for the classification of malicious webpages included static and dynamic heuristics [13], Honey client based approaches [28] and conventional machine learning techniques [29]. The conventional machine learning techniques used till date offered scope for improvement in accuracy, precision, and recall of classification results. The feature sets used were not comprehensive, leaving good scope for improvement. We have done an extensive feature analysis to recommend a set of features for the webpage classification task. Further, since deep learning technology has been successfully applied to many domains over the past few years, using deep learning for improving webpage classification results came as a natural extension. The work presented in the thesis uses both structured data (pre-processed feature sets) and unstructured data (raw web content) for analyzing and proposing suitable deep learning models for web security. Further details on research gaps are presented in

Chapter 4 for conventional machine learning based work, and in chapters 5 and 6 for deep learning based approaches.

- *Mitigating Threats Emanating from Android Hybrid Apps*: Till date, to the best of our knowledge, security architecture in hybrid apps [19] had not been analyzed in detail from the perspective of current web threats. Further, machine learning has not been utilized for such an analysis of hybrid apps. We have attempted to bridge this research gap in the thesis (further details on research gaps are presented in Chapter 7).

- *Privacy Preserving Solutions for Web Security using Cross-device Federated Learning*. Federated Learning (FL) is a new ML paradigm introduced by Google in 2017 [30]. Google uses FL for Android keyboard predictions [31]. In healthcare, FL has been applied for brain tumor segmentation without the data being moved out of the partnering hospital [32]. A machine/deep learning solution typically requires data from various sources/sites to be uploaded to a central server or a cloud where machine learning algorithms try to find patterns in the data. There are two major problems with this form of machine learning, called "Centralized Machine Learning" in literature. Firstly, the data needs to be communicated to the central server. In this era of Big Data, it can become a bottleneck. Second, and most important, is the fact that users' privacy is compromised when data leaves their device, machine, or organization. Most of us are not comfortable with sharing our data with anyone for any purpose. Federated learning addresses both these issues elegantly and has found ready applications in healthcare [33] and finance [34]. However, FL has not yet been explored in the field of Cybersecurity. In the thesis, we have exploited cross-device FL for solving web security related problems (refer to Chapter 8). Federated learning has spawned new research in Private/Edge AI, Secure Multiparty Communication (SMC), and Homomorphic Encryption (HE). FL is a single point failure system. If the FL server fails due to some hardware or communication problem, then the whole system fails. In the thesis, we have also proposed a Hierarchical Federated Learning (HFL) solution to address this issue. In addition to addressing the single

point of failure problem, HFL also gives added advantage of enhanced privacy and ability to examine local patterns.

1.4 Main Contributions

In the thesis, an attempt has been made to apply the latest developments in AI to solve some intriguing and challenging problems related to web security. Furthermore, the thesis has taken a holistic view of the web security challenges by considering various platforms, viz., computers, mobiles, etc., and has attempted to provide security solutions that are effective, deployable, and scalable. The main contributions of the thesis are summarized below:

- MalCrawler, a focused crawler for seeking and crawling malicious webpages, was proposed and developed. MalCrawler seeks more malicious webpages than a generic crawler and is designed to overcome evasive techniques utilized by malicious sites. MalCrawler has proven to be an effective tool for collecting data related to malicious webpages and malwares by cybersecurity firms and researchers.

- Conventional ML based models have been used as part of this thesis to carry out the classification of malicious webpages. Various attributes, which can be used for such classification, were analyzed as part of this research. The classification metrics surpassed the results of similar experiments. A Self Organising Map (SOM) [35] based analysis of the attributes was also carried out to identify various clusters in the dataset of malicious webpages.

- Deep learning based approach has been used to carry out the classification of malicious webpages. Deep learning was executed using both structured data (pre-processed and cleaned dataset with various attributes) and unstructured data (raw webpage content). The deep learning approach surpassed metrics of all similar ML attempts, including the conventional ML based approach used in this thesis.

- In the thesis web security on the mobile platform was also covered. A ML based analysis of Android hybrid apps was carried out, including

apps currently hosted on the Google Play store. Such a study helped in identifying vulnerable apps hosted on the store. Based on the machine learning study results, various corrective measures have been suggested to improve web security on the Android mobile platform. As part of this work, an Android app named 'WebView Tool' was developed and published on the Google Play store to understand better the hybrid app's WebView component and its exploitation by JavaScript-based injection attacks. Furthermore, another Android app named 'WebView Monitor' was designed and developed to monitor the working of hybrid apps on the Android platform. This app raises alerts during untrusted privilege escalation or when malicious JavaScript is downloaded.

- A FL based mobile web security solution was proposed and implemented. To the best of our knowledge, this is the first time that FL has been used for mobile security. The solution overcame the limitation of centralized ML based mobile security solutions, specifically with regard to privacy concerns of mobile users. Furthermore, a Hierarchical Federated Learning [36] (HFL) solution was proposed as part of the work, which allowed local patterns at appropriate levels of spatial granularity to be discovered along with the global patterns. Hierarchical federated learning also makes any FL system more robust by reducing the dependence on the main FL server. Overall, the proposed work on FL and HFL will pave the way for privacy-preserving mobile security solutions, with the added provisions for obtaining local/country wise patterns along with the global patterns

1.5 Thesis Organization

This thesis has been organized into three parts as brought out below:

- The first part, which includes Chapters 2 and 3, is related to data collection, pre-processing, and visualization for AI-based web security analysis on all platforms, including mobiles. MalCrawler, a special purpose focused crawler for seeking and crawling malicious webpages, has been described in Chapter 2. In Chapter 3, the pre-processing,

preliminary analysis, and visualization of the webpages and the hybrid apps dataset used in the thesis is presented. The experimental setup and code for reproducing these datasets is given in Appendix A and B respectively.

- The second part of the thesis proposes machine learning based analysis and solutions for webpages classification. This part comprises Chapters 4, 5, and 6. Chapter 4 uses conventional ML based approach, while Chapters 5 and 6 use deep learning based approach. Chapters 4 and 5 use structured data for classification, whereas Chapter 6 uses unstructured web content for classification.

- The third part of the thesis consists of research related to web security on the mobile platform. Chapters 7 and 8 comprise this part. Chapter 7 proposes a classification model using centralized ML to analyze hybrid apps on the Android platform. Chapter 8 presents a FL and HFL (distributed machine learning) based cross-device web security solution for the mobile platform.

- Chapter 9 concludes the thesis with recommendations/ suggestions and proposes scope for future work.

- Finally, to facilitate future research, a summary of all software and tools used for the research is given in Appendix C.