Automated Hardware Recon—A Novel Approach to Hardware Reconnaissance Process

Kalpesh Gupta, Aathira Dineshan, Amrita Nair, Jishnu Ganesh, T. Anjali, Padmamala Sriram, and J. Harikrishnan

Abstract Technology is growing at an exponential rate, and everything around us from shopping to instant messaging and emails, studies, etc. is getting connected to the Internet and becoming smarter. This enabled reconnaissance (or recon) is a collection of procedures and methods, including enumeration, foot-printing, that are used to covertly discover and acquire information about a target system. The protracted recon process requires utmost attention and precision when it comes to handling the device for inspection. There exists a high risk of tampering with the device while inspecting the interior, requiring the replacement of the device. With FCC ID or chip number extraction using optical character recognition, followed by double-checking with the dataset, the specifically designed web scrapers will help to scrape all the information required, including the vulnerabilities from the web, after which a brief report will be generated. Hence, our proposed system automates the process of reconnaissance, saving time and helps in avoiding risks to an extent.

Keywords Hardware recon · Reconnaissance · OCR · Dataset · Web scrapper · Beautiful Soup · Selenium · Hardware security · Datasheet

1 Introduction

The Internet is a crucial part of today's era, and from shopping to instant messaging and emails, academics, everything surrounding us is getting connected to the Internet and becoming smarter. This enables instant communication and interaction and provides simple access to information and services. To make these things smart, smart sensors, micro-controllers, and microprocessors are used in the devices. And

K. Gupta (✉) · A. Dineshan · A. Nair · J. Ganesh · T. Anjali · P. Sriram
Department of Computer Science and Engineering,
Amrita Vishwa Vidyapeetham, Amritapuri, India
e-mail: anjalit@am.amrita.edu

P. Sriram
e-mail: padmamala@am.amrita.edu

J. Harikrishnan
Cisco Systems, Bangalore, India

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 336, https://doi.org/10.1007/978-981-16-6723-7_20

with the growing use of such devices, protection of our information from getting misused is must. The software needs to be protected from malware, viruses, etc., and this is achieved by using anti-virus software. And with the increase in usage of smart devices, protecting the hardware from getting compromised becomes necessary too.

The hardware consists of various microprocessors and integrated circuits, and due to its complexity, it is hard to detect vulnerabilities and fix them as replacing hardware components can be a tedious task. And, it is difficult to find the source of the vulnerability, as there is a possibility of it being a manufacturing defect.

In the Cyber Security domain, pen-testing is a process of analyzing and assessing the "secureness" of a device by doing a series of simulated attacks on the device and looking for vulnerabilities. This tells the pen-testers the flaws of the device and gives them an idea of the possible attacks. In the hardware domain, pen-testers use a process called reconnaissance.

Reconnaissance (or recon in short) is a collection of methods and processes, including scanning and enumeration, that are used to secretly uncover and collect knowledge about the target systems. There are two kinds of recon processes—active and passive reconnaissance [1]. Active reconnaissance is a type of computer intrusion in which an attacker interfaces with the targeted device to collect information about vulnerabilities [2]. Passive reconnaissance is the method of gathering information about the intended victim of a malicious hack without the target knowing what is going on [3].

Hardware reconnaissance is a process composed of various steps like device disassembly, looking at various components, port debugging like JTAG/SWD, chip identification, and information extraction from its datasheet. Before invading the target, a full understanding of the device is required, even though it is a black box during the process of pen-testing or security analysis. The recon phase helps to identify multiple components of the system so that one's attacks can be targeted toward what one knows, including the vulnerabilities found when learning about the system.

Every electrical and electronic product with a working frequency of more than 9KHz needs to be FCC certified. The FCC stands for Federal Communications Commission. FCC regulations are designed to reduce electromagnetic interference, manage and control a range of radio frequencies to protect the normal work of telecommunications networks and electrical products [4]. Wireless devices or products with wireless transmission frequency are assigned an FCC ID. It is a unique identifier assigned to a device registered with the United States Federal Communications Commission [5]. Using the FCC ID, one can obtain photographs of the device, user manuals for the device, etc. [6].

Another source of information for getting details about the chips and microcontroller is the manufacturer's Web sites where the information such as the datasheets, operating information is available, but it is in a very scattered manner and requires a lot of manual intervention to search the required information.

Web scraping is the practice of gathering organized web data in an automated manner. It is also called extracting web info. Any of the key applications of web scraping includes, among many others, pricing tracking, price intelligence, news monitoring, market survey, and sentiment analysis. It is also called extracting data [7–9].

A Python library to pull data from HTML and XML files is Beautiful Soup. It functions to include idiomatic ways to browse, find, and alter the parser tree using the parser. It usually saves hours or days of work for programmers. It also automatically converts incoming docs to unicode and outgoing documents to UTF-8 [10].

Selenium is a cross-platform framework based in JavaScript, Python, C#, Ruby whose development was first started by ThoughtWorks, by a person named Jason Huggins when he was building a testing application for an internal expenses and time application [11]. Selenium is predominantly used for extracting data from dynamic web pages and building automated functional tests for testing web applications.

This work proposes an easy-to-use web application-based tool which automates the hardware recon process by automatically extracting the FCC ID or the chip number of the chip from the device image and generates a brief report containing the details of the past exploits, vulnerabilities in the device, operating temperature, datasheet links, etc.

The main contribution of our work is that we are proposing a novel approach for the recon process and an extensive pre-compiled dataset. The proposed approach combines the optical character recognition with the recon process and lookup in this extensive dataset to automate the recon process, saving time and avoiding risks to an extent.

The rest of this paper is structured as follows. Some of the related existing works and approaches toward web scraping and text extraction from images are in the second section. The third section describes our proposed approach including the details of the dataset that is compiled from a variety of web resources to fasten up the recon process. The further sections describe the report generation process, benefits of the proposed system, conclusion and future works.

2 Related Works

Ray Smith describes the steps involved in the text extraction using Tesseract, including processing, recognition, and classification of the image, to extract the text character by character [12]. Nguyen et al. perform a statistical analysis the possible errors caused by the optical character reader using four different datasets [13]. Payel Roy et al. compare different algorithms by comparing the adaptive threshold values using Correlation and Structural Similarity Index (SSIM) calculations [14].

The algorithm proposed by S. Chaudhari et al. uses the web scraper tool Scrapy and MongoDb in the application. The application stores the recipe name, ingredients, and the URL of the recipe in the database, collected through web scraping beforehand [15]. Shinde Santaji Krishna and Joshi Shashank Dattatraya presented a page-level data extraction system that extracts web page schema from template generated web pages automatically [16]. Ahmad Pouramini and Shahram Nasiri proposed a tool that generates web scrapers to extract data items from the web page. They have tried to stimulate the way humans look at web pages and have used textual anchors to create patterns for the target data regions [17]. Sanya Goel et al. propose a PHP-based web application which is able to crawl through useful information from the schools' Web site and provide aid to parents in the Delhi NCR region [18]. S. Thivaharan. et al. compared the popular web scraping libraries such as Beautiful Soup, LXml, and RegEx in terms of response time (best, average, and worst cases) and accuracy [19].

3 Proposed Solution

Our suggested approach starts with getting an input of the image of the device from the user of the device or chip under study for the vulnerabilities and exploits. Upon receiving the input, the extraction phase starts to extract the text from the input image and process it to get the FCC ID or the chip number using the OCR engine. The extracted chip number is searched in the extensive dataset to get the resource links and other details. Then, the required information is web scrapped from those resource links, while for the FCC ID, the information is directly web scrapped from the order to generate a brief report for the user containing the details of the past exploits, vulnerabilities, and other important device information. The proposed approach is summarized in Algorithms 1 and 2.

Algorithm 1: Hardware recon using the FCC ID

```
Input  : An image of the device containing FCC ID
Output : Report containing datasheets, internal, and external images of the device

1. Preprocessing of the input image;
2. Passing the processed image to the OCR engine to extract the FCC ID ;
3. Extracted FCC ID is appended to the URL “www.fccid.io\” ;
4. Web scrape all the datasheets, internal, and external images of the device;
5. Generate a combined report;
```

The diagram displayed in Fig. 1 represents the proposed solution using FCC ID and chip number.

3.1 OCR Engine

OCR or optical character recognition is a text recognition system that recognizes and extracts text from digital documents and images. It is widely used in AI, machine

```
Algorithm 2: Hardware recon using chip number
Input  : An image of the chip/device containing chip number
Output : Summarized report of the chip along with CVE links

1. Preprocessing of the input image;
2. Passing the processed image to the OCR engine to extract the chip number;
3. Extracted chip number is searched in the managed dataset;
4. if chip number is found then
    i. Retrieve the resource links and other information from the database ;
    ii. Web scrape device information from the Web site ;
    iii. Generate a report from the retrieved data ;
else
    Generate error message for the chip number not found.
end
```

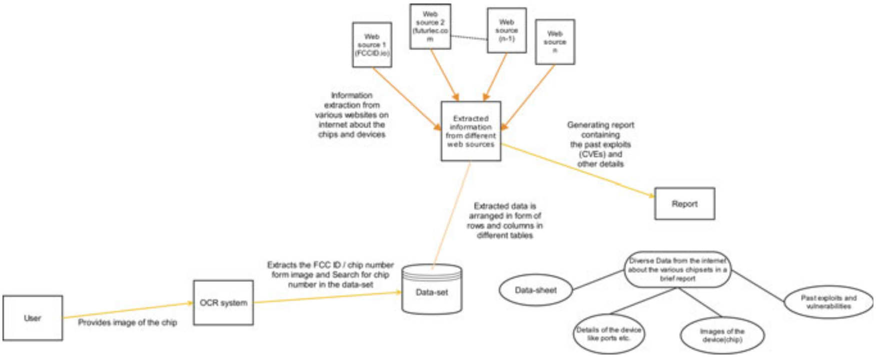

Web source 1 (FCCID.io)
Web source 2 (futurlec.com
Web source (n-1)
Web source n
Information extraction from various websites on internet about the chips and devices
Extracted information from different web sources
Generating report containing the past exploits (CVEs) and other details
Report
Extracted data is arranged in form of rows and columns in different tables
Extracts the FCC ID / chip number form image and Search for chip number in the data-set
Data-set
User
Provides image of the chip
OCR system
Data-sheet
Diverse Data from the internet about the various chipsets in a brief report
Past exploits and vulnerabilities
Details of the device like ports etc.
Images of the device(chip)

Fig. 1 Block diagram of the proposed solution for automated hardware recon

learning, robotics, IoT, banking, and health care [20]. One of the most popular and commonly used OCR engines is Tesseract. It is open sourced and identifies a wide range of languages. PyTesseract, which is the OCR tool for Python, is used for our system.

In the proposed system, the OCR engine is used for FCC ID extraction or the micro-controller chip number extraction of the scanned images, based on the user's need.

3.1.1 FCC ID Extraction

Typically, the device details are found in the form of printed text. Thus, in this case, the prepossessing involves converting the image into grayscale and increasing the contrast so as to make the text to appear more prominent [21].

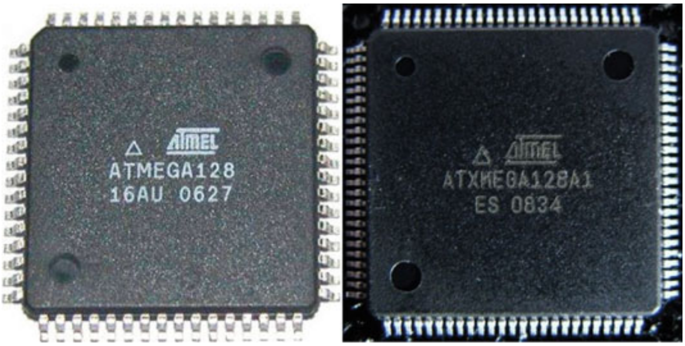


Fig. 2 Comparison between new and old micro-controller texts

3.1.2 Chip Number Extraction

In the case of micro-controllers, the text is not printed, rather etc.hed on the mounting package of the integrated circuit. As a consequence over time, the chip number may lose its clarity (Fig. 2).

As a result, the OCR may or may not recognize the letters accurately. After several trials and errors, two different preprocessing methods which yielded better results concerning the system were proposed.

1. Edge Detection: In this case, the noise is removed, and edges of the letters are detected via Canny edge detection. This method works well in cases where the text engravings are not distinguished and the micro-controller surface is more smooth.
2. Thresholding : In cases where the letters are not defined or the edges are not very clear, the image is dilated to accentuate the letters, and this is followed by thresholding.

So, the image is passed through the above two methods, and the processed image is then fed to Tesseract to obtain the chip number. The output from the methods is then cross-checked with a predefined array which consists of common prefixes of the different micro-controller families. If a match exists, a lookup of the particular chip number is done against the dataset to obtain the corresponding resource and CVE links.

3.2 Dataset Construction

Although there are thousands of micro-controller or communication modules used today worldwide, the most commonly used can be found in various applications around us like that of Arduino (Atmega series), Microchip's PIC series (PIC 32,

Table 1 Sample of records in the dataset

Chip number	Resource link	CVE1	CVE2	CVE3
ATSAMA5D21	⟨Manufacturer site⟩	⟨CVE link 1⟩	⟨CVE link 2⟩	⟨CVE link 3⟩
ATSAM3U2C	⟨Manufacturer site⟩	⟨CVE link 1⟩	⟨CVE link 2⟩	⟨CVE link 3⟩
:	:	:	:	:
:	:	:	:	:

18, 16, 12), Microchip's ATSAM series, ESP32 series, TI's C2000 series, TI's SimpleLink Series, Ti's MSP430 series, Arm's Cortex series,Arm's Mali GPUs.

Arduino micro-controller is used in variety of smart IoT-based systems today due to its cheaper cost and high usability. It is also used in the domestic air quality monitoring and gas leak detection systems [22]. It is also used in the IoT-based smart bins [23].

Hence, we have compiled an extensive dataset consisting of the various resource links of these commonly used micro-controllers and chip set from various sources on Internet. These resource links point to the Web sites from where we can get much information about the chip sets like the types of interface supported, chip family, DRAM interface types, micro-controller images, number of pins, whether supports GPU or not, and other chip specific information like I2C, SSC, datasheets, and other important resources about the chip set.

Besides resource links, the common vulnerabilities and exposures (CVE) links for the micro-controller chips present in the dataset (wherever available) were also included, from where one can get to know about the vulnerabilities in the smart devices and take appropriate actions to resolve them. If there are multiple CVE links available for some chips, they were included so that most appropriate information can be obtained in the brief report generated by the proposed system.

This compiled dataset is composed of the details of over 1000 most commonly used chip sets which will aid us in our proposed automated hardware recon process.

The different family of micro-controllers is arranged across different sheets to enable faster search operation. A sample of the type of record in the dataset in a sheet can be seen in Table 1.

3.3 Resource Link Lookup

Once the chip number from the OCR engine component of the proposed approach is obtained, one can lookup for it in our extensive dataset (which we mentioned above) among the families in which the chip belongs which can found using the initial part of the chip number. Once the chip number is found in the dataset, one can retrieve the resource link from there to start the information extraction process. Also, one can retrieve the CVE links for that particular chip from the dataset.

If FCC ID is obtained from the chip, one can directly move to the data extraction part (FCC ID part) of the proposed approach.

3.4 Data Extraction

The resource links that are retrieved in the previous stage (in the case of chip number) are used to scrape the important information from the Internet. Python modules such as Beautiful Soup and Selenium framework are used in our proposed system to web scrape important information about the micro-controller chip from the different Web sites on the Internet as denoted by the resource links.

The information from the Web sites is extracted using the X-paths of the sections where the information is located on the web page using the Selenium framework. The extracted information includes the port types, number of pins, supported interfaces that can be useful for pen-testers and hardware researchers. .

For FCC ID, the information about the chip can be web scraped from fccid.io, including the internal and external images of the device under study.

4 Report Generation

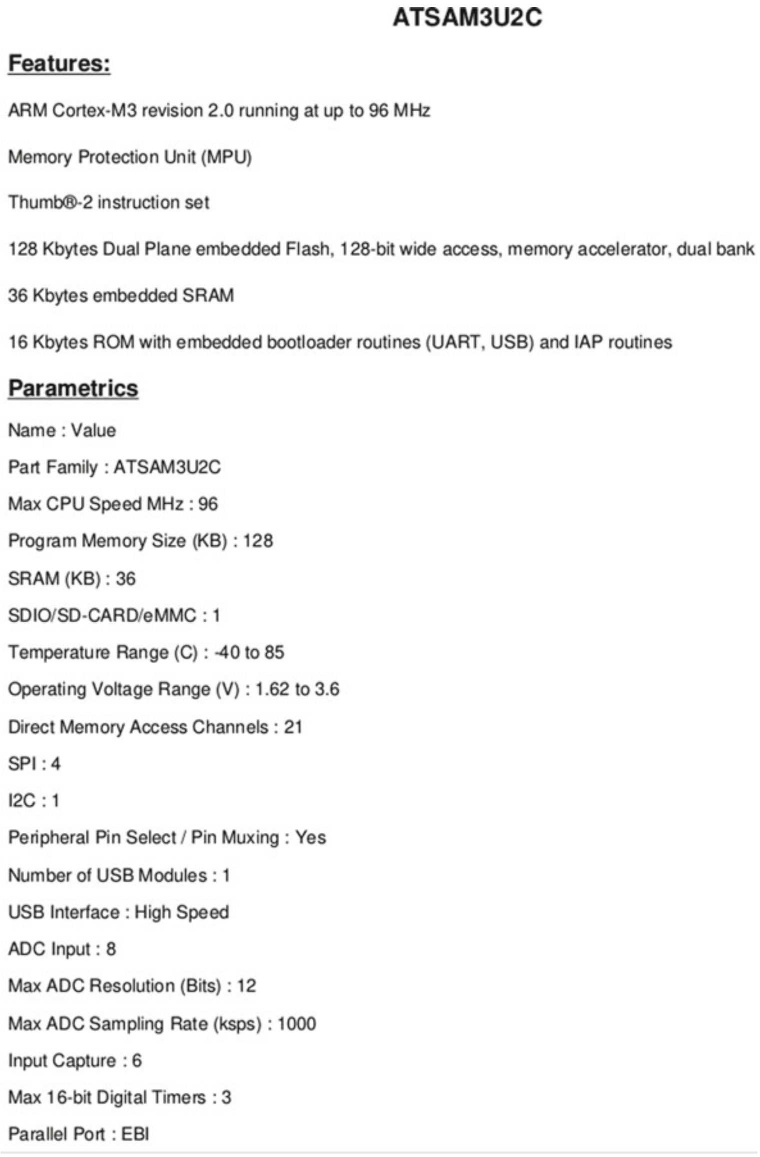
The extracted information from the previous stages is compiled in a brief report in PDF format. Besides the information about the micro-controller chip and device under study, it will also include the CVE links that are retrieved from the dataset which can help the pen-testers and hardware researchers to know about the past vulnerabilities and exploits found in the micro-controller chip.

This entire process can be easily completed by the user by utilizing our web application by simply providing the image of the device, and our proposed system will perform the recon process and generate a brief report for the user.

5 Experimental Results

Hence, the proposed automated system can generate a brief report consisting of micro-controller chip features, para-metrics, CVE, and datasheet links scraped from the Internet which can aid the pen-testers and hardware researchers in reconnaissance process. This will help them to save time, focusing on other important work to enhance hardware security.

A sample snip of the report is shown in Figs. 3, 4 and 5.



ATSAM3U2C

Features:

ARM Cortex-M3 revision 2.0 running at up to 96 MHz

Memory Protection Unit (MPU)

Thumb®-2 instruction set

128 Kbytes Dual Plane embedded Flash, 128-bit wide access, memory accelerator, dual bank

36 Kbytes embedded SRAM

16 Kbytes ROM with embedded bootloader routines (UART, USB) and IAP routines

Parametrics

Name : Value

Part Family : ATSAM3U2C

Max CPU Speed MHz : 96

Program Memory Size (KB) : 128

SRAM (KB) : 36

SDIO/SD-CARD/eMMC : 1

Temperature Range (C) : -40 to 85

Operating Voltage Range (V) : 1.62 to 3.6

Direct Memory Access Channels : 21

SPI : 4

I2C : 1

Peripheral Pin Select / Pin Muxing : Yes

Number of USB Modules : 1

USB Interface : High Speed

ADC Input : 8

Max ADC Resolution (Bits) : 12

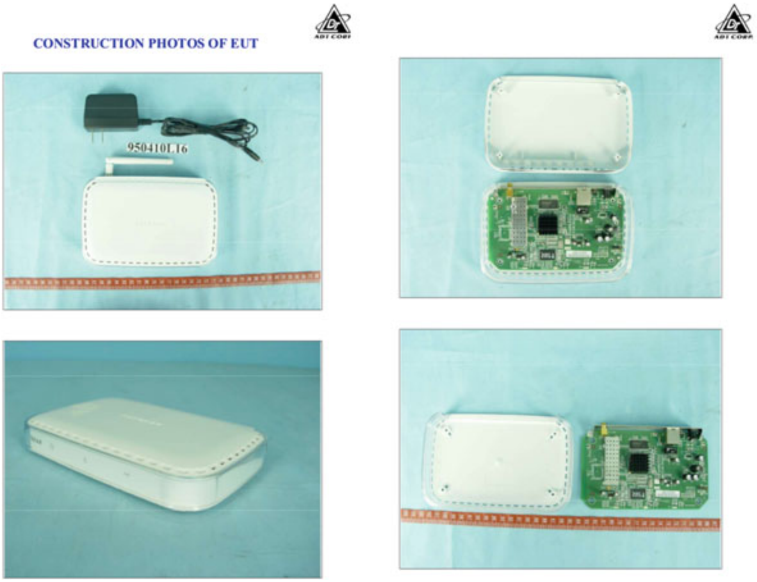
Max ADC Sampling Rate (ksps) : 1000

Input Capture : 6

Max 16-bit Digital Timers : 3

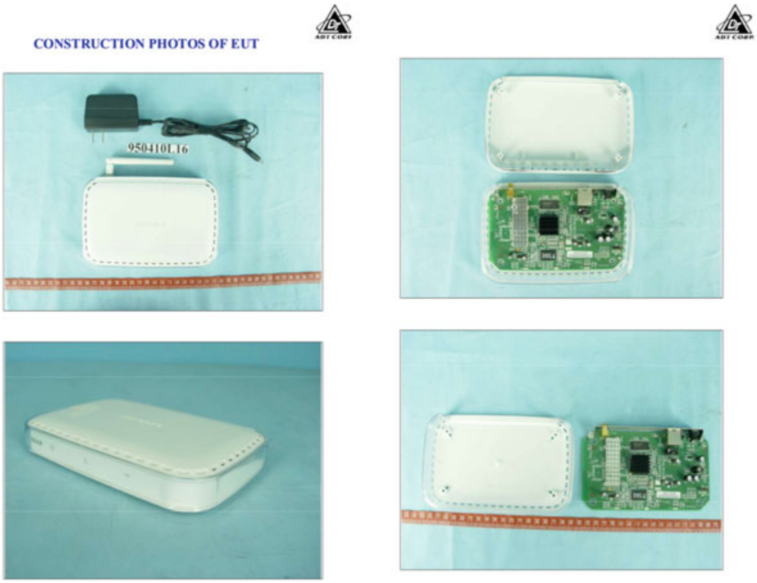
Parallel Port : EBI

Fig. 3 Sample snip from the report (Part 1)



CONSTRUCTION PHOTOS OF EUT

Fig. 4 Sample snip from the report (Part 2)



CONSTRUCTION PHOTOS OF EUT

Fig. 5 Sample snip from the FCC ID report

6 Advantages

Pursuing the vision of our university, we directed our research efforts toward compassion-driven research. This proposed novel solution will be able to reduce the man-hours and efforts put into the pen-testing and reconnaissance processes. It will also help in eliminating human errors that might slip in during the manual process, in turn helping to reduce the amount of e-wastes produced due to discarding of the tampered devices. Toxic substances present in these e-wastes are lethal both to the environment and the beings; thus, this proposed automated approach will contribute to lowering it to a certain extend.

7 Conclusion and Future Work

Hence, a novel approach is proposed that automates the traditional hardware recon process using the combined vigor of OCR, our compiled dataset, and web scrapping to generate a brief report containing important information about the micro-controller chip which may act as an aid for pen-testers and hardware researchers. And, all of these can be controlled using an easy-to-use web application.

The work can be further enhanced by developing a better and efficient OCR for refining the process of FCC ID and chip number extraction. Dataset expansion by including new micro-controllers and the development of an algorithm to collect recent CVEs can contribute to make the application versatile. Research on adaptable web scrapers can help in replacing the customized web scrapers, thus saving time in developing those for newly included micro-controllers.

Acknowledgements We would like to utilize this section for expressing our gratitude toward our source of inspiration, Mata Amritanandamayi Devi, affectionately known as Amma, Chancellor of Amrita Vishwa Vidyapeetham who has devoted her life to selfless service to the society. We would also like to extend our sincere gratitude to our mentors and teachers who have always helped us whenever we faced any difficulties and motivated us to continue this work. Also, not to skip, thanks to all the people who have directly and indirectly aided us in this work.

References

1. Ethical Hacking—Reconnaissance—Tutorialspoint, in Tutorialspoint.com. https://www.tutorialspoint.com/ethical_hacking/ethical_hacking_reconnaissance.htm. Accessed 8 May 2021
2. Active Versus Passive Reconnaissance—ASM, Rockville, Maryland, in ASM, Rockville, Maryland. https://asmed.com/active-vs-passive-reconnaissance/. Accessed 8 May 2021
3. Passive Versus Active Reconnaissance, in *Medium*. https://medium.com/@jharve08/passive-vs-active-reconnaissance-c2974913237f. Accessed 8 May 2021

4. What is the FCC-ID authentication? Huawei Enterprise Support Community, in *Huawei Enterprise Support Community*. https://forum.huawei.com/enterprise/en/what-is-the-fcc-id-authentication/thread/588944-883. Accessed 8 May 2021
5. Fccid, in Fccid.io. https://fccid.io/. Accessed 8 May 2021
6. A. Dineshan, G. Gokul Krishna, J.L.A. varshini, J. Ganesh, T. Anjali, J. Harikrishnan, Hardware security reconnaissance application using FCC ID lookup and computer vision, in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 526–529. https://doi.org/10.1109/ICCSP48568.2020.9182318
7. What is web scraping and how does it work? Zyte.com, in Zyte (formerly Scrapinghub) #1 Web Scraping Service. https://www.scrapinghub.com/what-is-web-scraping/. Accessed 8 May 2021
8. T. Anjali, T.R. Krishnaprasad, P. Jayakumar, A novel sentiment classification of product reviews using Levenshtein distance. Int. Conf. Commun. Signal Process. (ICCSP) **2020**, 0507–0511 (2020). https://doi.org/10.1109/ICCSP48568.2020.9182198
9. A.C. Jishag, et al., Automated review analyzing system using sentiment analysis, in *Ambient Communications and Computer Systems. Advances in Intelligent Systems and Computing*, vol. 904, eds. by Y.C. Hu, S. Tiwari, K. Mishra, M. Trivedi (Springer, Singapore, 2019). https://doi.org/10.1007/978-981-13-5934-7_30
10. L. Richardson, Beautiful Soup: We called him Tortoise because he taught us, in Crummy.com. https://www.crummy.com/software/BeautifulSoup/. Accessed 8 May 2021
11. Selenium history, in Selenium.dev. https://www.selenium.dev/history/ Accessed 8 May 2021
12. R. Smith, An overview of the Tesseract OCR engine, in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Parana, 2007, pp. 629–633. https://doi.org/10.1109/ICDAR.2007.4376991
13. T. Nguyen, A. Jatowt, M. Coustaty, N. Nguyen, A. Doucet, Deep statistical analysis of OCR errors for effective post-OCR processing, in *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, Champaign, IL, USA, 2019, pp. 29–38. https://doi.org/10.1109/JCDL.2019.00015
14. P. Roy, S. Dutta, N. Dey, G. Dey, S. Chakraborty and R. Ray, Adaptive thresholding: a comparative study, in *2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kanyakumari, 2014, pp. 1182–1186. https://doi.org/10.1109/ICCICCT.2014.6993140
15. S. Chaudhari, R. Aparna, V.G. Tekkur, G.L. Pavan, S.R. Karki, Ingredient/recipe algorithm using web mining and web scraping for smart chef, in *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, Bangalore, India, 2020, pp. 1–4. https://doi.org/10.1109/CONECCT50063.2020.9198450
16. S.S. Krishna, J.S. Dattatraya, Schema inference and data extraction from templatized Web pages, in *2015 International Conference on Pervasive Computing (ICPC)*, Pune, 2015, pp. 1–6. https://doi.org/10.1109/PERVASIVE.2015.7087084
17. A. Pouramini, S. Nasiri, Web data extraction using textual anchors, in *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, Tehran, 2015, pp. 1124–1129. https://doi.org/10.1109/KBEI.2015.7436204
18. S. Goel, M. Bansal, A.K. Srivastava, N. Arora, Web crawling-based search engine using python, in *3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. Coimbatore, India, 2019, pp. 436–438 (2019). https://doi.org/10.1109/ICECA.2019.8821866
19. S. Thivaharan, G. Srivatsun, S. Sarathambekai, A survey on python libraries used for social media content scraping, in *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2020, pp. 361–366. https://doi.org/10.1109/ICOSEC49089.2020.9215357
20. D. Kanteti, D.V.S. Srikar, T.K. Ramesh, Intelligent smart parking algorithm. International Conference On Smart Technologies For Smart Nation (SmartTechCon) **2017**, 1018–1022 (2017). https://doi.org/10.1109/SmartTechCon.2017.8358524

21. V. Bharath, N.S. Rani, A font style classification system for English OCR, in *2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, pp. 1–5. https://doi.org/10.1109/I2C2.2017.8321962
22. K. Gupta, G. Gokul Krishna, T. Anjali, An IoT based system for domestic air quality monitoring and cooking gas leak detection for a safer home, in *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 0705–0709. https://doi.org/10.1109/ICCSP48568.2020.9182051
23. A. Praveen, R. Radhika, M.U. Rammohan, D. Sidharth, S. Ambat, T. Anjali, IoT based Smart Bin: a Swachh-Bharat initiative. International Conference on Electronics and Sustainable Communication Systems (ICESC) **2020**, 783–786 (2020). https://doi.org/10.1109/ICESC48915.2020.9155626