Real-World Antivirus Evaluation Methodology: Applying Modern Criteria for Assessing Antivirus Functionality

Youngrak Ryu*
KAIST Cyber Security Research Center
Daejeon, South Korea
yrryu@kaist.ac.kr

Kangsik Shin
KAIST Cyber Security Research Center
Daejeon, South Korea
ksshin90@kaist.ac.kr

Jeongho Lee
KAIST Cyber Security Research Center
Daejeon, South Korea
ddanzit@kaist.ac.kr

Dong-Jae Jung
KAIST Cyber Security Research Center
Daejeon, South Korea
jjp1018@kaist.ac.kr

Ho-Mook Cho‡
KAIST Cyber Security Research Center
Daejeon, South Korea
chmook79@kaist.ac.kr

***Abstract*—This study proposes a new approach to antivirus testing environments and testing methods that differ from the criteria developed by conventional antivirus test institutes. Nine known antiviruses were tested under environments that resemble real-world use scenarios with advanced evaluation criteria. Through the test results, we evaluated the performance of each antivirus and identified the advantages and disadvantages of antiviruses in different situations. This study argues that the existing antivirus evaluation criteria are inappropriate for evaluating modern antiviruses and proposes a highly accurate method for measuring antivirus.**

***Index Terms*—antivirus, method, criteria, malware, real-world**

I. Introduction

Since the outbreak of the COVID-19 pandemic in 2020, we have suffered from the fear of the relentless virus and have fought against this virulent epidemic. The U.S. federal government spent over $30 billion on COVID-19 vaccines [1]. The amount financial lost by the COVID-19 pandemic in the U.S. by the end of 2023 will be about $14 trillion [2]. Similarly, in the cyber world, computer virus is constantly attacking cyber assets to steal data and damage computers and systems in cyberspace. The global economic loss from cybercrime will reach nearly $1 trillion annually [3]. Typical end-users protect their computers to defend against the computer virus with antivirus (AV) designed to prevent, detect, and remove intruding computer viruses. Thus, the AV is one of the most essential software when setting up a new computing system. According to Security.org's report, 85% among the surveyed 1,003 adult users in the U.S. have installed AV on their computers [4].

It would be decent for users to pick the best AV based on their situations. However, in reality, most users and businesses blindly use well-known AVs. In some countries, government-made guidelines separating and recommending trustworthy AVs are published to encourage using certain products. This is because there are too many AVs, and because most of them get perfect ratings from testing institutes. Speaking of the testing institutes, they conduct malware protection tests or real-world protection tests regularly, and they give most AV vendors perfect scores. The test results show that most AV SWs detect and block over 90% of malware [5], and we thought the results are abnormal and suspicious. Therefore, we decided to conduct our own tests of various AV SWs in real-world conditions to verify if the existing results are accuracy or not.

In this paper, we establish detailed evaluation methodologies suitable for evaluating AV and apply them to nine off-the-shelf AVs to analyze functionality and performance.

This paper is organized as follows. Section 2 describes related research, Section 3 introduces a designed test environment and establishes the AV evaluation criteria, Section 4 evaluates the AVs with the pre-designed evaluation criteria, and finally, Section 5 concludes the achievements of this study and gives implications for future work.

II. Related Work

A. Antivirus Software

In recent years, malicious software, known as computer virus, have evolved and diversified their infection methods and symptoms, resulting in a number of different types of malware, adware, and spyware. Thus, in order to combat the diversity of evolving malwares, AVs had also been evolving. Recent AVs have different analysis capabilities and detection engines to find and remediate different types of malicious software. The essential features and functions of an AV are

- Periodic and automatic detection pattern updates

This work was developed with the support of Global Cybersecurity Research grant (Project ID: 1711177169) from the Ministry of Science and ICT, Republic of Korea.

979-8-3503-3094-6/24/$31.00 ©2024 IEEE

- Real-time surveillance detection
- Manual and regular virus scanning

Moreover, modern AVs have become a multi-security software including built-in privacy protection, network security such as firewall and VPN, and system protection that allows log file access and control storing medium functions.

B. Antivirus Test Institutions

AV test institutes conduct various tests under independent criteria to evaluate the security performance, reliability, and usability of AV, and issue certificates if it satisfies their criteria. They publish the results in the media and on their blog and report reviews of the tested software. AV vendors regularly have their products evaluated by the test institutes to evaluate performance and earn certifications objectively. Most major vendors achieve positive ratings from the test institutes and use the results for advertising their products and influence customers to trust them. The test institutes that conduct these evaluations and issue certificates are described below.

- **ICSA Labs** [6]: Founded in 1991, ICSA Labs is an independent test institute in the United States that certifies various information protection products, such as cryptographic equipment, intrusion prevention and detection security systems, and AVs. It is a globally trusted certification institute that conducts government-certified security standards testing and cryptographic module verification(CMVP) evaluations for the U.S. Department of Defense. AV evaluations are conducted with tens of thousands of malware samples in three types: In-the-Wild, Common Infectors, and Zoo, and require 100% accurate detection and no false positives to obtain certification, making it a demanding and prestigious certification organization.
- **Comparitech** [7]: Comparitech is a UK-based institute founded in 2015 that tests various products, including VPNs, AVs, network monitoring tools, firewalls, and more. It provides detailed reviews of AVs, including their strengths, weaknesses, pricing, comparisons between different AVs, AV ranking, and test results. The tests are conducted in a sandbox environment in various ways, including detection, remediation, auto-renewal policy, system impact, and primary and advanced functions. For detection, tests are conducted using EICAR test malware and real-world malware, and evaluations are conducted using real-time scans and full system scans.
- **AV-Comparatives** [8]: An independent security product testing institute based in Austria that conducts monthly or quarterly tests of security products for personal and business use on a diverse range of topics. In particular, it conducts tests under real-world conditions and collects samples of malicious sites using its own crawling system. To be certified, a product must satisfy the organization's 13 requirements for reliability and stability. AV-Comparatives' Real-World Protection Test methodology has been recognized with awards such as the Constantinus Award from the Austrian government and the Cluster Award from Standortagentur Tirol.
- **AV-TEST** [9]: AV-TEST is a global security product performance evaluation institute based in Germany that tests and ranks different AV products. Products are tested in three categories: protection (detection/cure), usability, and performance, and can earn up to 6 points each, with a total of 10 points or more and at least 1 point in each category to be certified. It conducts various types of tests on OS platforms such as Android, MacOS, and Windows, individuals, enterprises, and IoT devices and announces the test results on its website.
- **Virus Bulletin** [10]: A private British security research institute that publishes a magazine specializing in malware. It has been publishing its magazine since 1989, and since 2014 has only published standalone articles on its website. Its certification program, VB100, has operated since 1998 and is a certification test for Windows endpoint security solutions. It is tested on a virtual machine with 1,000 to 2,000 recent malware and 100,000 legitimate applications that have had real-world infection or discovery reports in at least two regions, and the evaluation period is approximately one month. Certified products and results will be posted on the homepage in the latest order.
- **MRG Effitas** [11]: Started as an online forum in 2009, the UK-based independent performance rating institute has created its own levels to evaluate AVs. They develop their own malicious apps that reflect the latest malware trends for the test. AV is tested with 500 samples of legitimate applications and various malware, including exploits, ransomware, botnets, and adware, with a malware detection rate of at least 99%. If they pass all quarterly tests, they are awarded an Effitas award, and the results are published on a website.
- **SE Labs** [12]: A British independent testing institute that conducts various AV tests and provides consumer reports on security products for individuals and businesses. The tests are conducted on Windows PCs, isolated from other target systems using VLANs, and the sample malware is weighted to be widespread. Tests run quarterly and are categorized into enterprise, small business, and consumer products.

C. Existing Antivirus Test Principle

Since the threat of malicious malware, such as viruses, has grown, the importance of AV has also increased. Therefore, there are many different AVs on the market, and it is essential to study the criteria to evaluate them. Dunham published a journal, ”Evaluating Antivirus Software: Which is Best?” and considered essential factors for AV test are cost, system requirements, interface and leadership, performance, scanning options, removal and recovery options, support, and compatibility [13]. Willems et al. introduces the Anti-Malware Testing Standards Organization (AMTSO) in his book, ”Cyberdanger.”

The AMTSO proposed nine essential principles for the AV test.

1) Testing must not endanger the public.
2) Testing must not be biased.
3) Testing should be reasonably open and transparent.
4) The effectiveness and performance of AV must be measured in a balanced way.
5) Testers must take reasonable care to validate whether test samples or cases have been accurately classified as malicious, innocent, or invalid.
6) Testing methodology must be consistent with the testing purpose.
7) The conclusions of a test must be based on the test results.
8) Test results should be statistically valid.
9) Vendors, testers, and publishers must have an active contact point for testing-related correspondence.

These nine principles ensure that AV testing is consistent, useful, and efficient and that a test product that satisfies these criteria fulfills its testing purpose [14].

III. OUR TESTING METHODOLOGY

The existing evaluation criteria for AV needs to evaluate software performance adequately. To improve the existing evaluation methods to meet the evaluation, this paper comprehensively classifies common evaluation criteria and functions based on the Korean ”Software Technical Evaluation Criteria Guidelines,” ISO/IEC 25010 [15], ISO/IEC 25020 [16], ISO/IEC 25023 [17], ISO/IEC 25041 [18], and the manuals of each AV, and establishes evaluation criteria by simplifying the evaluation items.

The evaluation criteria are divided into six major categories:

1) functionality to evaluate AV performance, such as malware detection and speed
2) resource efficiency to check excessive resource use of the system
3) reliability to assess the stability of the AV
4) usability to evaluate the overall UI/UX and convenience
5) add-ons for determining the provision of additional features
6) vendor support to update and troubleshoot

This study focuses on the functionality of the six evaluation categories, the fundamental AV evaluation. The results are analyzed quantitatively. The remaining criteria are evaluated qualitatively, and results are described.

A. Selection of AV Software

In the test, we focus on AVs using Microsoft Windows OS. There are more than thirty AVs on the market with more performance variation and additional features. Among them, the software for the test was selected based on the following reasons.

1) High market share in South Korea and the world
2) Product with unique engine
3) For Windows operating systems
4) Includes a free AV

Based on these criteria, we selected nine AVs for testing, as shown in Table 1.

TABLE I
ANTIVIRUS SOFTWARE FOR TESTING

Antivirus Software	Version
ESTsecurity Alyac	5.1.22
AhnLab V3 Internet Security	9.0
Avast Premium Security	23.3.6058
Kaspersky AV	21.3.10
McAfee Total Protection	16.6.161
MS Defender	4.18.2303
ESET Nod32 AV	16.1.14
Norton AV	22.23.3
TrendMicro Internet Security	17.7.1827

TABLE II
TEST PC SPECIFICATION

	Specification
CPU	Intel i5-12400
GPU	On-board Intel UHD Graphics 730
Memory	DDR4 16GB
Storage	SSD 500GB
OS	Windows 10 Pro (64bit)

B. Test Methods

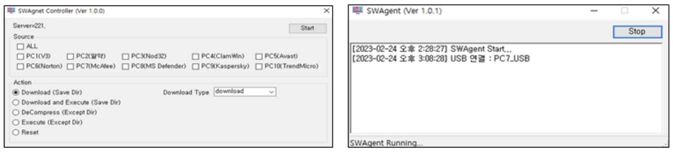
The malware to be used in the evaluation was selected by considering the type, appearance, and characteristics of malware collected from malware DB sites such as MalShare [19] and VirusShare [20], including a massive pool of malware provided by VirusTotal [21] for research purposes and malware collected from South Korean public organizations, security companies, and our own crawling system. Depending on the evaluation scenario, we used five malware samples for each scenario.

1) 100 malware categorized by extension (exe, pdf, hwp, et al.)
2) 20,330 malware selected by randomly sampling a large amount of malware
3) 151 malware in the last three months collected from South Korean public organizations and our crawling system
4) 50 malware packed with commercial packing tools
5) 25 executable malware

Email phishing is a technique that involves sending emails to organizations or public institutes with document-type malware attached. Therefore, we included document-type malware in our malware sample and added HWP file malware, a popular document format in South Korea. The ability to detect such region-specific malware is also one of the evaluation measures in this test. We conducted testing and performance evaluation by introducing testing methods not used by other certification authorities, such as USB flash drives, large amounts of malware, and recently collected malware.

C. Test System Architecture

We prepared the test computers with the specifications shown in Table 2 and preinstalled applications such as messaging, word processing, and VOD on the test PCs to mimic a real-world environment for a typical user closely. In addition to the test PCs, a server PC was set up to distribute the malware to each PC and store test logs. To avoid human mediation, we developed a command execution application to execute each scenario, as shown in Figure 1. Using the controller and agent, we tested all test PCs simultaneously, including malware deployment, execution, file extraction, and rebooting. The test logs of the commands executed by each PC and the files and actions transferred are stored in the DB.



SWAgnet Controller (Ver 1.0.0)

Server=221.

Source
□ ALL
□ PC1(V3) □ PC2(알약) □ PC3(Nod32) □ PC4(ClamWin) □ PC5(Avast)
□ PC6(Norton) □ PC7(McAfee) □ PC8(MS Defender) □ PC9(Kaspersky) □ PC10(TrendMicro)

Action
● Download (Save Dir)
○ Download and Execute (Save Dir)
○ DeCompress (Except Dir)
○ Execute (Except Dir)
○ Reset

Download Type: download

SWAgent (Ver 1.0.1)

```
[2023-02-24 오후 2:28:27] SWAgent Start...
[2023-02-24 오후 3:08:28] USB 연결 : PC7_USB
```

SWAgent Running...

Fig. 1. Controller and agent applicaton.

The test system architecture consists of a controller server and DB, test PCs, and a tester to collect and analyze test results. The system was implemented with a dual firewall to prevent possible malware leakage. Figure 2. shows an overview of the system architecture.

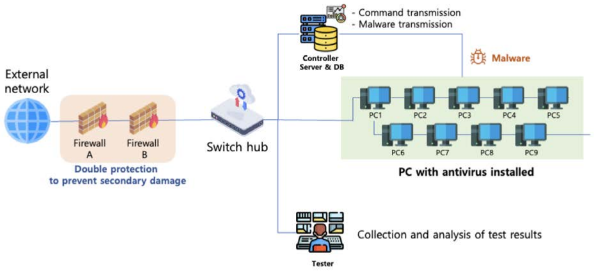


Fig. 2. Overview of the test system architecture.

IV. EVALUATION

In this chapter, we describe the results of the tests based on preconfigured scenarios and explain the evaluation of each test. The test took place over two months and was conducted in a separate lab with no external access to prevent interference from the outside environment. The AVs subject to the test were anonymized using alphabet letters from "A" to "I".

A. Categorized 100 Malware Test

We tested the real-time detection capabilities of 100 malware organized by file extension. The malware consisted of 17 exe files, 17 xlsx files, 17 pdf files, 17 hwp files, 16 pptx files and 16 docx files, all determined malicious by VirusTotal [21]. This experiment aims to see if AV products can correctly detect malware in real time when transmitted to a computer over a network. Among the AVs, H had the best detections, with 75. It performed exceptionally well in detecting exe and pdf files. Test 1-2 in Figure 3. uses the same malware as in Test 1-1, but instead of transmitting the files in real-time from the server, the detection results were measured after extracting the compressed files. Once again, H had the best detection performance with 70 detections. Some AVs were unable to detect any malware in real-time.

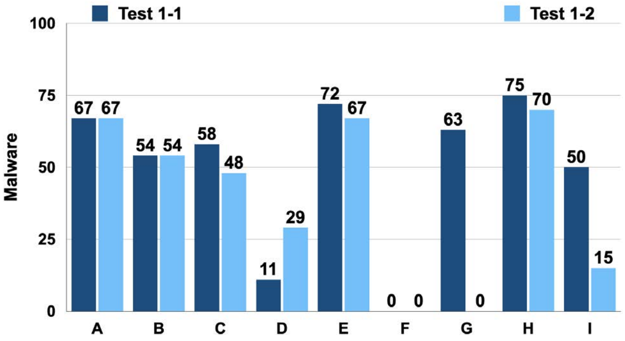


Fig. 3. The result of the categorized 100 malware test.

B. 25 Executable Malware Test

The second test is the real-time detection accuracy of 25 executable malware. This experiment aims to evaluate the accuracy of behavior-based malware detection. The malware was executed sequentially on all test PCs via a controller application, and the detection results were checked. After executing the malware, we rebooted the PCs to restore them to their pre-infection state, as this could affect the system. The experiment results showed that F detected all 25 behavior-based malware with a 100% detection rate, which is the best performance.

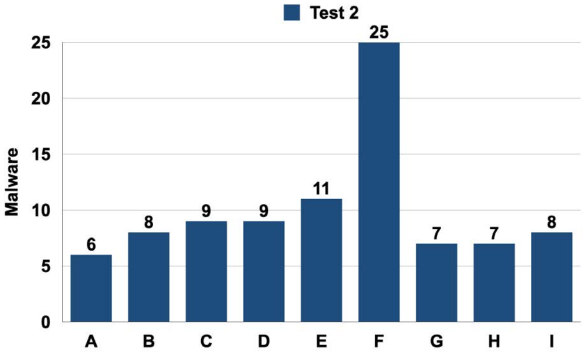


Fig. 4. The result of the 25 executable malware test.

C. 20,330 malware selected by randomly sampling a large amount of malware

In this test, we measured the accuracy of detecting a large set of 20,330 malware. It was conducted in two ways

1) External disk scanning using a USB flash drive containing malware

2) Real-time detection by extracting the malware and testing the detection accuracy

In Test 3-1, we used a USB flash drive to detect malware using the external disk scan function. As a result of the test, F detected 17,439 malware (85% detection rate), showing the best detection accuracy. C showed 17,247 malware (85% detection rate), showing excellent results with a close difference from F.

In Test 3-2, the controller application was used to test the real-time detection capabilities of the test PCs by simultaneously extracting the malware archive. In this test, B performed the best, detecting 14,871 malware (73% detection rate). However, the other AV tests failed to detect malware or had meager detection rates. It is believed that they were not seen because they may not be in the detection zone for the duration of the tests.

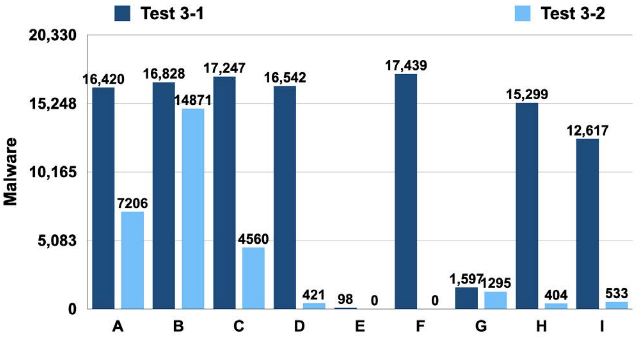


Fig. 5. The result of the large set of 20,330 malware test.

D. 151 malware in the last three months collected from South Korean public organizations and our crawling system

It evaluates the latest 151 unknown malware collected within the last three months. The malware set was organized by our own crawling system and malware provided by the South Korean government. The malware were sent to each test PC simultaneously by a controller application using the same method as in Test 1. The results showed that C detected 108 malware with a detection rate of 71.5%, indicating the best accuracy. Overall, the results were good, but D and F did not perform as well.

E. 50 malware packed with commercial packing tools

This test measures the real-time detection accuracy of packed malware utilizing a commercial packing tool. We packed 50 malware samples using Themida from Oreans Technologies [22], a commercial packing tool. Test 5-2 is a control group that detected unpacked malware for comparison. The test methodology was the same as Tests 1 and 4, using a controller application to send to each test PC simultaneously. As shown in Test 5-1 of Figure 7, the packed malware detection test showed the best results, with E detecting 37 (74%) of the packed malware, compared to 33 (66%) of the unpacked malware. G had the best detection accuracy in the control group test with 49 (98%) but significantly lower detections in the packed malware detection test with 13 (26%). Most AVs show lower detection rates when detecting packed malware. However, only E showed an increase in detection rates.

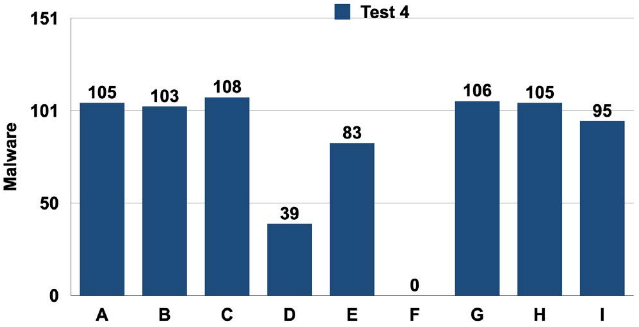


Fig. 6. The result of the latest 151 unknown malware collected test.

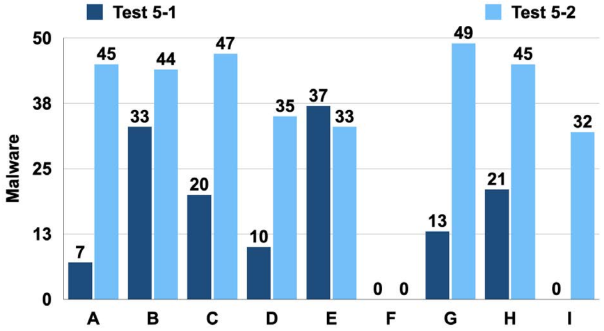


Fig. 7. The result of the packed 50 malware test.

F. Detection Speed Evaluation

In Test 6, we measured the time for a deep scan performed on a USB flash drive with a large amount of malware to measure the correlation between the results from Test 3 and the detection time. We measured when the USB flash drive was inserted and terminated for this detection speed evaluation to determine the overall scan time. C took the longest scan at 6 hours, 12 minutes, and 13 seconds. A took an estimated 11 minutes and 35 seconds. Since C detected 17,247 malware, 85% of the detection and second-best detection rates, we investigated the correlation between the detection rate and the time spent. Figure 8. shows the correlation between scan time and detection rate. E and G, which had low detection rates, took 15 and 16 minutes, while A, B, and I detected over 10,000 malware in a low scan time. We could not find a correlation between detection time and detection rate in this experiment. We found that detection time depends on the engine and detection method used by the AV.

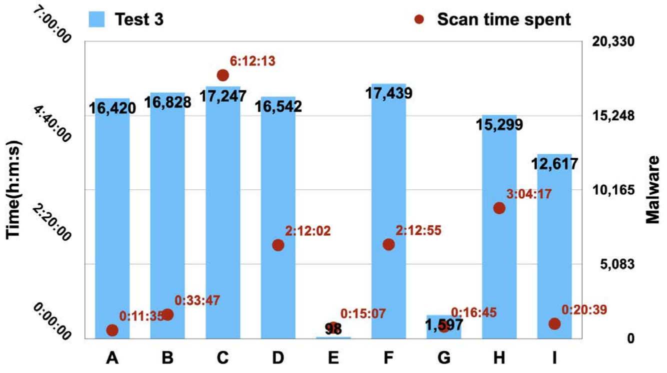


Fig. 8. The correlation between scan time and detection rate.

G. Analysis and Discussion

With the results obtained from the measurements, we analyze the results of AV detection. We tested a variety of situations, including real-time detection, behavior-based malware detection, massive malware detection, external storage detection, recent malware detection, and packed malware detection. We found that certain AVs performed strongest in only behavior-based situations, while others performed well in all situations. We found that the detection rate varies depending on the engine and detection method used by each product. When we looked at the correlation between detection accuracy and detection speed, we did not find a significant relationship between speed and accuracy.

We measured AVs' reliability by detecting large amounts of malware. While detecting large amounts of malware, none of the nine products crashed during our tests, and none crashed during stress tests that involved running with other applications, indicating that all AV products were highly compliant. Regarding usability, most of the commercial products supported Mac OS, Linux, and mobile OS. They offered paid and free business models with support plans, giving consumers a wide range of choices. In addition to malware detection and remediation, they provided a rich set of add-ons, including network security, registry cleaning, and privacy protection. All AV vendors supported regular updates to defend against the frequent emergence of new malware and supported multiple languages.

V. RESULT & FURTHER WORK

This study was motivated by doubts about the results of several AVs achieving over 90% correct detection rates and earning "recommended" badges from various AV testing institutes. We set up a real-world test environment and designed five scenarios to measure detection results in different situations. We refined the evaluation measures, developed the test environment, and conducted tests according to each scenario. The test results were quite different from the results of existing testing institutes. In particular, we found that AV F had high detection accuracy only in behavior-based malware detection and very low detection accuracy in other situations. In this way, we identified the characteristics of each AV engine. By evaluating nine AVs, we validated the evaluation criteria and detailed evaluation measures and characterized the features and performance of different AVs.

In future studies, we will regularly evaluate new AV based on the established evaluation criteria. In particular, we will improve the accuracy of detection speed by evaluating only detected files when evaluating detection speed, and we will enhance the evaluation criteria to include the evaluation of enterprise versions that contain most of the features on the management server. Furthermore, based on this test's evaluation criteria and items, we will expand to a feature and performance verification study for mobile AVs.

References

[1] J. Kates, C. Cox, and J. Michaud, "How much could covid-19 vaccines cost the u.s. after commercialization?" KFF, 12 2022. [Online]. Available: https://www.kff.org/coronavirus-covid-19/issue-brief/how-much-could-covid-19-vaccines-cost-the-u-s-after-commercialization/

[2] T. Gann, "The hidden costs of cybercrime on government," McAfee Blogs, 12 2020. [Online]. Available: https://www.mcafee.com/blogs/other-blogs/executive-perspectives/the-hidden-costs-of-cybercrime-on-government/

[3] T. Walmsley, A. Rose, R. John, D. Wei, J. P. Hlávka, J. Machado, and K. Byrd, "Macroeconomic consequences of the covid-19 pandemic," *Economic Modelling*, vol. 120, p. 106147, 2023.

[4] S. Team, "Personal antivirus consumer usage, adoption & shopping study: 2021," Security.org, 02 2023. [Online]. Available: https://www.security.org/antivirus/antivirus-consumer-report-annual/

[5] AV-Comparatives, "Real-world protection test february-may 2022," AV-Comparatives, 06 2022. [Online]. Available: https://www.av-comparatives.org/tests/real-world-protection-test-february-may-2022/

[6] ICSA labs. [Online]. Available: https://www.icsalabs.com

[7] "Comparitech - tech researched, compared and rated," Comparitech. [Online]. Available: https://www.comparitech.com

[8] AV-Comparatives. [Online]. Available: https://www.av-comparatives.org

[9] "Av-test — antivirus & security software & antimalware reviews," www.av-test.org. [Online]. Available: https://www.av-test.org

[10] "Virus bulletin," www.virusbulletin.com. [Online]. Available: https://www.virusbulletin.com

[11] "Mrg effitas - world-leading, independent it security testing," MRG Effitas. [Online]. Available: https://www.mrg-effitas.com

[12] "It security testing by se labs: Next-gen, full attack chain testing," SE Labs. [Online]. Available: https://selabs.uk

[13] K. Dunham, "Evaluating anti-virus software: Which is best?" *Inf. Secur. J. A Glob. Perspect.*, vol. 12, no. 3, pp. 17–28, 2003.

[14] E. Willems, *Cyberdanger : understanding and guarding against cybercrime*. Springer, 2019.

[15] I. O. F. STANDARDIZATION, "Iso/iec 25010: Systems and software engineering-systems and software quality requirements and evaluation (square)," 2011.

[16] ——, "Iso/iec 25020: Systems and software engineering-systems and software quality requirements and evaluation (square) - quality measurement framework," 2019.

[17] ——, "Iso/iec 25023: Systems and software engineering: Systems and software quality requirements and evaluation - measurement of system and software product quality," 2016.

[18] ——, "Iso/iec 25041: Systems and software engineering: Systems and software quality requirements and evaluation - evaluation guide for developers, acquirers and independent evaluators," 2012.

[19] "Malshare," www.malshare.com. [Online]. Available: https://www.malshare.com

[20] "Virusshare.com," www.virusshare.com. [Online]. Available: https://www.virusshare.com

[21] VirusTotal, "Virustotal," 2019. [Online]. Available: https://www.virustotal.com

[22] "Oreans technologies : Software security defined." www.oreans.com. [Online]. Available: https://www.oreans.com/themida.php