INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Investigation on File Carving Tool Methodologies Using Scenario Based Image Creation

Hepi Suthar[1]*, **Priyanka Sharma**[1]

1 AI & Cyber Security Department, School of Information Technology, Rashtriya Raksha University, Gandhinagar, Gujarat, India

OPEN ACCESS

Received: 07-04-2023

Accepted: 19-12-2023

Published: 12-01-2024

Citation: Suthar H, Sharma P (2024) An Investigation on File Carving Tool Methodologies Using Scenario Based Image Creation. Indian Journal of Science and Technology 17(3): 215-227. https://doi.org/10.17485/IJST/v17i3.808

* **Corresponding author**.

hepisuthar@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The objective of this study is to develop and validate carving techniques and tools for recovering fragmented files in digital forensics, using various data sets (Data recovery done from various SSD for investigation environment) to verify effectiveness and contribute to the advancement of the field. **Methods:** The research method used in this study involves the development and validation of carving techniques and tools that can effectively retrieve data from fragmented files. The study uses various data with the use of scenario based from different research organizations to verify the effectiveness of the developed carving techniques and tools. Here this study based on foremost tool, the study also creates 16 pictures for tool verification depending on situations and uses a well-known commercial carving tool, Foremost, to demonstrate the developed image carving pace and precision of each media. The research method is focused on functional verification and creating solutions for digital forensics. **Findings:** The research findings of this study indicate that the data recovery (from solid state drive) of fragmented files is a significant challenge for file carving (Tools - Scalpel, Bulk Extractor, Foremost, and Photorec) in digital forensics. The study highlights the need for developing and validating carving techniques and tools that can effectively retrieve data from fragmented files. The data sets (Data recovery done from various SSD for investigation environment) from different research organizations were used to verify the effectiveness of the developed carving techniques and tools. The study creates 16 pictures for tool verification depending on situations and uses Foremost (Forensic Tool), a well-known commercial carving tool, to demonstrate the developed image carving pace and precision of each media. The study also found that various carving techniques and tools are constantly being created to get around these restrictions. However, the current data sets (Data recovery done from various SSD for investigation environment) are less useful for validating tools due to their constrained environmental circumstances. Therefore, the study recommends the use of more realistic data sets (Data recovery done from various SSD for investigation environment) to validate carving techniques and tools. Overall, the study's findings contribute to the advancement of digital forensics by providing a more efficient and reliable solution for recovering data from fragmented files. The study's results can be used to improve the accuracy and effectiveness of file carving techniques and tools, thus enhancing digital forensic investigations. **Novelty:** The novelty of this research lies in developing and validating carving techniques and tools for effectively retrieving data from fragmented files in digital forensics. Like various scenario tested with forensic tools likes Scalpel, Bulk Extractor, Foremost, and Photorec. And tested based on various file signature (document, audio, video, email, and archive). Here file carving technique major with this linearly stored file carving performance, deleted files, file carving performance stored non-linearly, non-linearly, Master Boot Recorder, GUID Partition Table, Hard Disk Drive, and Solid State Drive.

Keywords: SSD; Digital Forensics; File Carving; HDD; Data Recovery

1 Introduction

In the ever-evolving landscape of digital forensics, the recovery of valuable data from digital devices remains a challenging endeavor. A critical component of this process is the utilization of file carving tools, which play a pivotal role in the extraction of fragmented or hidden information from storage media. These tools must continually adapt to the dynamic nature of data storage and retrieval, necessitating an investigation into their methodologies. This paper, titled ”An Investigation on File Carving Tool Methodologies Using Scenario-Based Image Creation,” delves into the intricacies of file carving methodologies with a particular focus on the creation and validation of scenario-based images. By doing so, it seeks to provide valuable insights into the effectiveness, accuracy, and adaptability of file carving tools in digital forensic investigations. In this introduction, we will outline the significance of file carving tools in the field of digital forensics, underscore the importance of scenario-based image creation, and offer a brief overview of the paper’s objectives and methodologies(1).

- **Significance of File Carving Tools** - Digital forensics is an essential discipline that aids in the investigation of criminal activities, cybersecurity incidents, and data recovery. It relies heavily on the recovery of information from various storage devices, such as hard drives, solid-state drives, and mobile devices, often in cases where data has been intentionally or unintentionally deleted, concealed, or damaged. File carving tools play a pivotal role in this process by identifying and extracting files and fragments of data from unallocated space or storage media. These tools are indispensable for piecing together evidence in investigations, recovering lost information, and uncovering hidden or deleted files. The accuracy and efficiency of these tools are of paramount importance to digital forensic investigators.
- **Scenario-Based Image Creation** - One of the challenges in digital forensics is the validation and testing of file carving tools in real-world scenarios. Traditional testing environments often fall short in replicating the complex, multifaceted situations encountered in forensic investigations. To bridge this gap, scenario-based image creation becomes an invaluable technique. Scenario-based image creation involves generating test data that mimics actual forensic scenarios. This method encompasses a wide range of situations, such as data storage on different media types, data fragmentation, and varying storage conditions. By utilizing scenario-based images, digital forensic investigators can assess the performance and accuracy of file carving tools under conditions that closely resemble real-world cases(2).

The primary objective of this paper is to investigate file carving tool methodologies, with a specific focus on their effectiveness when applied to scenario-based images. Through a systematic approach, we aim to evaluate the adaptability of these tools to diverse forensic scenarios and their ability to retrieve data accurately. To achieve this, our research methodology includes the creation of scenario-based images representing a variety of digital forensic situations. These images will serve as a robust testing ground for file carving tools. Subsequently, we will conduct in-depth analyses of the performance, precision, and limitations of these tools in scenarios that closely emulate those encountered by digital forensic investigators.

2 Methodology

Digital devices leave a considerable range of records about individuals as data. Therefore, the importance of digital forensics is increasing in recent litigation, focusing on digital evidence. As an important point in a survey targeting digital devices, the issue of how much deleted data can be recovered became an issue, and forensic hardware and software tools for file recovery appeared to meet these requirements. Techniques for file recovery are generally divided into metadata-based file recovery and carving based file recovery. Carving-based file recovery is a recovery method used when a file is deleted or the file system is formatted, and the file metadata information is changed to other data or disappears, making it impossible to access the file, and traces of evidence destruction can be found. Therefore, it is an essential element in digital forensics. However, in file carving, performance differences may occur depending on the functional or structural differences of the storage medium may be saved and stored. As such, the file carving function of tools may differ in various cases, so it is necessary to verify the tools in an environment that reflects realistic situations from various scenarios, and it is necessary to supplement the limitations of the tools according to the verification results(2). However, it is difficult to verify the reliability of the tool because images for verification of file carving reflecting various situations are not provided in existing studies and projects. In this paper, Hard Disk drive and Solid State Drive, Master Boot Record and GUID Partition Table development environments and disks with un-fragmented files (S1), disks formatted from S1 (S2), disks with fragmented files stored (S3), and 16 realistic file carving tools that reflect 4 scenarios for a disk formatted in S3 (S4) are presented for verification. Section 2 introduces the importance of carving and projects and studies that provide an image of existing tool verification and presents its limitations. Section 3 describes the effect on carving tools according to the storage medium, and Section 4 describes the file fragmentation and deletion scenarios. And section 5 describes the results of verifying the foremost tool using the presented carving verification image.

2.1 The need for file carving tool verification

Recovering deleted files is one of the important steps in investigating digital devices. In general, when a file is simply deleted, the metadata area is not deleted, but a specific flag value is changed and information about the file such as name, size, and allocation location is maintained. It is possible to restore the original state. However, if the file system does not exist or is damaged, or even more, if the storage medium is formatted or the metadata is overwritten with another value, it is difficult to recover the file because the location of the actual file data is unknown. There are limitations to the recovery technique based on it. This limitation is overcome by using a database of headers and footers for file formats to search for files regardless of the file system of the disk image (without reference to offset or sector). It is also possible Therefore, in the case of file carving, it is possible to recover a file by using the file system structure or even if the metadata of the file system is destroyed, so a relatively large number of deleted files can be recovered. It is effective. In response to these requirements, file carving software tools of various techniques have appeared, as well as embedded file carving functions in existing computer forensic tools. In terms of external functions, self-proclaimed file carving tools were scattered, but as reliability of the tool function became a problem, verification of the tool function became necessary. Various studies on file carving tool verification have been conducted according to necessity, and images for testing are provided[3].

2.2 Related Studies

The CFTT (Computer Forensics Tool Testing) project of the National Institute of Standards and Technology (NIST) in the United States defines the requirements for tools used in digital forensics and verifies each tool. Through the CFReDS (Computer Forensic Reference Data Sets) project, methods and procedures were established and a test environment was established. For the file carving test, we provide graphical, document, compressed, audio, and video files and six different levels of fragmentation scenarios. As each attribute file is composed of one file, the total number of data sets (Data recovery done from various SSD for investigation environment) for each image is less than 10, and the image size is also small, with a maximum of 50 MB. In addition, the generated image is artificially created and does not have a file system structure, which may reduce reliability in tool verification. DFTTI (Digital Forensics Tool Testing Images) is a project carried out in 2003 to close the gap between CFTT's research on tool validation conducted by public institutions and research on validation techniques conducted by private organizations. For file carving tool verification, it targets two file systems, FAT32 and EXT2, and saves USB flash drives to mkfs. vfat, mkfs. It provides an image formatted through ext2[4]. The number of test files consists of 15 or less, which makes the dataset very small to validate and evaluate the tool. RDC (Real Data Corpus) closely imitates the data actually found by purchasing discarded devices from secondary markets around the world without deleting or deleting data from storage devices used by general users and extracting the data. Create and provide data sets (Data recovery done from various SSD for investigation environment) that The NPS test disk image is publicly available as a disk image set created for testing computer forensic tools. Among them, in the image data set taken with a Canon digital camera, which can verify the carving of basic files and fragmented files, some of them are fragmented, but the files are limited to JPG. The DFTTI Basic Data Carving Test No. 1 (11-carve-fat.dd), the DFRWS2006 Forensics Challenge dataset (dfrws-2006-challenge.img. Using a single baseline carving dataset (bcds.raw) - Baseline Carving Data Set, the carving performance of EnCase, FTK, WinHex, PhotoRec, Scalpel, and Foremost tools was compared and verified[5].

Table 1. Related work difference vs. proposed scenarios

Data Related work	Proposed Method	Variation
CFTT	Artificial images up to 50 Mega Byte (Non File system) Archive, Audio, Documnent, Graphic, Videofiles (Total 31) ● Non fragmented files ● Sequential fragmentation ● Non Sequential fragmentation ● Missing fragments ● Nested (sub files) Files ● Braided files	-100Mega Byte Natural OS image (OS Windows 10). -New Technology File-system. -Hard Disk Drive (HDD) & Solid State Drive (SSD). -Master Boot Recorder & GPT.
DFTTI	● 32MB USB Flash drive images ● FAT32 & EXT2 ● Audio, Document, Graphic, Video file (Total 25) ● mkfs.vfat & mkfs. ext2 ● Non fragmented files ● Sequential fragmentation	-Different types of file extension likewise Audio, Document, Graphic, Video, Email files, Archive (Total 241). -USB (Universal Serial Bus) booting format. -Files are not fragmented and the disks not formatted.

Continued on next page

Table 1 continued

RDC	• Flash memory card 32Mega Byte SD card image • New Technology File System • Total JPG files (Total51) • Non fragmented files • Sequential fragmentation • Non sequential fragmentation	-The secondary storage disk is format. -Files are fragmented without disk format. -The disk is formatted after files are fragmented.

Here found many fragmented and stored files by analyzing the drives obtained from the secondary market. As various image development studies for verifying file carving tools are continuously being conducted as in existing studies, it is very important to verify the importance of file carving and the reliability of the tools that perform it. Carving tools should be made large enough to detect performance improvements because small performance improvements have a direct impact on file recovery. This is an excellent way to begin the day. In addition, although such verification should be mainly evaluated experimentally, tools tested from limited images developed from existing studies show good verification results, but when put into actual investigations, more complex cases are encountered, and the performance results of the tools may vary, so the functionality is unclear. Table 1 of this paper an overview of the differences between the presented scenario and the previously presented scenario is presented[6].

2.3 File Carving Tools

According to various techniques, there are various tools that provide forensic tools with carving functions or only file carving functions. Because the file types or techniques provided by these tools are different, the file carving results for each tool in the same environment are different. Well-known open source file carving tools include Scalpel, Bulk Extractor, Foremost, and Photorec. Bulk extractor can carve zip and exif files, but most of them are used for carving character strings, and Photorec is used for carving graphic files. Foremost and Scalpel are file carving tools that support various files such as documents, audio, video, compression, and graphics. In this paper, the foremost v1.5.7 tool is used for tool verification in an environment composed of data sets (Data recovery done from various SSD for investigation environment) of various file types. The tool is widely known as a Linux platform open-source file carving tool for data recovery and was initially used to recover files using headers, footers, and data structures[6]. Image files can be created with DD, Safe back, Encase, etc., or image files can be directly carved into the drive. The header, footer, and data structure designate the files to be carved from the configuration (con-fig) file, so the specific file to be restored by entering the header and footer information of a job, you can select a file that is not supported by entering the attribute. Foremost was also used to develop a carving tool called Scalpel, and Scalpel currently shares the code with Foremost but uses an optimized method to reduce unnecessary memory-to-memory copy and disk I/O, and has relatively good performance. However, since Scalpel has more diverse header/footer signatures and data structures for file carving than Foremost, which is set by default, the number of files that can be carved increases unnecessarily[7]. There are limitations to the experimental environment in the study. Therefore, we present the verified result from the image developed through Foremost.

3 Results and Discussion

Image development environment/Proposed Method

3.1.1 Disk Structural Variations (MBR vs GPT)

The Microsoft Windows operating system provides two architectures for partitioning a disk drive into usable areas to store data. The difference between the architectures of the two architectures, MBR used in BIOS (Basic Input/output System) systems and GPT used in EFI (Extensible Firmware Interface) systems, is that depending on how the mapping is tracked, each method shows structural differences in disk management. In the MBR disk method, 3 primary partitions and 1 extended partition can be created, and the extended partition can be divided into multiple logical partitions. However, the number of addresses that can be recognized on the disk is, and the maximum size is 2TB. Due to the structural limitations of the MBR disk method, it is possible to create a maximum of 128 partitions per disk in the new disk format GPT, and the maximum volume of the volume has been extended to 18EB (Exabyte). A subset of that specification also includes a GUID (Globally Unique Identification) partition table or GPT header to replace the DOS/MBR partition table and contains backup data in the last space of the disk, so take advantage of it if files are deleted. It may be easier to perform file recovery than the MBR method. Therefore, if the disk recognition method is identified as GPT before performing file carving, it may be useful to analyze the backup data in advance. It should be different. It seems that if you use the recovery data area in the same way as in the carving tool, you will be able to carve a lot of files with more accuracy. Therefore, for the verification of the file carving tool in the structural difference of these disks, the corresponding scenario is included[8].

3.1.2 Disk Hardware Differences (HDD vs SSD)

Both HDD and SSD have recently been widely used as system drives. HDD, as the read/write head passes through the magnetic substrate, the data bits are sorted as 0, 1, or all. These sets of data bits together form bytes and are usually grouped into sectors (usually 512 bytes). However, unlike HDDs, SSDs write data electrically rather than magnetically, and when a delete command is given, data is completely deleted through the TRIM function rather than overwritten, and data is written in empty blocks. It is difficult to recover deleted files. The wear-leveling function of the SSD has a limited number of program-delete (P/E) cycles, so the controller executes writes so that the P/E cycle is evenly distributed for the entire block. For example, SSDs store data in pages of different sizes[9]. These pages are then grouped into erasable blocks, which are zoned together based on their physical address. Data is not sequentially written to the page but is striped across the erase blocks and managed by the wear-leveling controller. When the data stored on the disk is modified, the wear-leveling controller moves the entire block to a new location and then schedules the original block to be deleted. Unlike the HDD, it cannot be overwritten. This means that SSD users have no control over where data is written, and data is fragmented in a more complex way. As such, it may be difficult to recover data compared to HDDs in SSDs that do not have a saving pattern, as data is basically stored in the block with the least writes. In the real world, because different storage media are used, it is necessary to verify the reliability of the carving function in various environments.

Table 2. HDD vs. SSD [10]

Features	SSD	HDD
Mechanism	NAND NOR Flash Memory	Magnetic rotating platters
Capacity	Up to 1TB (notebooks) Up to 4TB (desktops)	Up to 2TB (notebooks) Up to 10TB (desktops)
Durability	Shock-resistant	Fragile
Power consumption	Average 2W	Average 10W
Endurance	MTBF > 2 million hours	MTBF < 700,000 Hours
Noise	None	Present
Operating System Boot-Time	Average of 10–15 seconds	Average of 30–40 seconds
File Opening Speed	30% faster than HDD	Slower than SSD
Speed	>200 MB/s	50–120 MB/s
Vibration	No Vibration	Moving parts cause vibration
Affected by Magnetism	No effect	Can erase data
Full Drive Encryption	Supported	Supported
Cost	Costly	Cheap compare to SSD
Heating	less	High
Size	Compact	Large

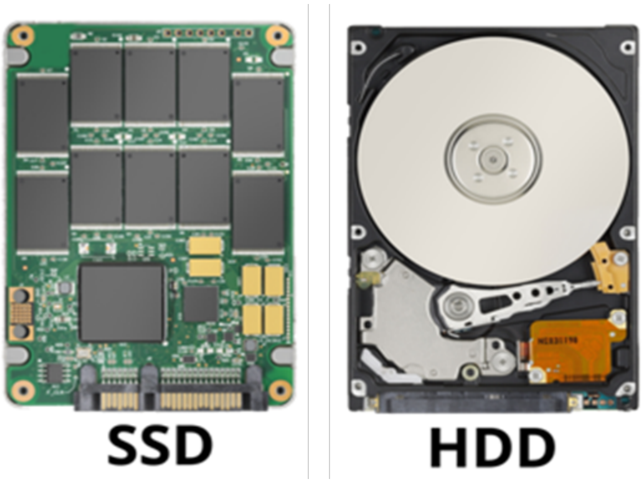


Fig 1. SSD and HDD Hardware Structure

3.2 Image development scenario/Experimental Work

The entire test image for the file carving test reflects the various cases that can occur in the four storage media environments using HDD, SSD, MBR, and GPT. The PC environment is the same as the Windows 10 operating system and the NTFS file system of 100 GB size. In Windows 10, the Trim function is enabled by default, and the Trim command is processed whenever a file is deleted from the SSD. So, for the scenario using SSD, it is after Trim works. In addition, when the data set from the four storage media is stored without fragmentation (S1-to measure the linearly stored file carving performance), when the storage medium is formatted in the environment of S1 (S2- Deleted files) to measure the carving performance), when the data set is fragmented and stored in the storage medium (S3-to measure the file carving performance stored non-linearly), and when the storage medium is formatted in the environment of S4 (S4-non-linearly) to gauge effectiveness of files after being erased & saved, it consists of four scenarios. In the cases of S1 and S3, since the storage medium is not formatted, the file carving tool can be used for the purpose of copying and restoring data using the structure of the file system[11].

Table 3. Validation Methodologies for file carving

Sr No	Object with Senario
S1 (linearly stored file carving performance)	The disk is not formatted, and the files are not fragmented. • To evaluate how well linearly stored carving files operate. • To assess the effectiveness of file carving within the file system.
S2 (Deleted files)	To assess the speed of carving deleted data, the disk is formatted in S1 (linearly stored file carving performance).
S3 (file carving performance stored non-linearly)	Without a disk format, files are fragmented. • To evaluate how well carving files stored non-linearly perform. • To evaluate the file system's ability to do file carving.
S4 (non-linearly)	The seconary storage is formatted in S3 (file carving performance stored non-linearly). • To assess the performace of file carving using the H file system.
Master Boot Recorder	To assess how well carving files operate in the MBR (master boot record) structure.
GUID Partition Table	To determine whether the utility has the capacity to carve files into a GUID partation table structure and recognize GPT.
Hard Disk Drive	To test carving performance in a relatively straightforward manner for lost and fragmented data.
Solid State Drive	To evaluate TRIM function ability to o complicated carving on fragmented files and deleted data.

In above Table 3, provides an overview of the descriptions and objectives of the four scenarios for image development and the objectives of Master Boot Recorder, GPT, Hard Disk Drive, and Solid State Drive, presented in sections 3.1 and 3.2. As a result, 16 images for the verification of the carving tool are created from 4 storage media and 4 scenarios.

3.2.1 File fragmentation

Carving fragmented files in digital forensics is becoming an important part. This is because, in recent forensic investigations, important files are more likely to be fragmented than other types of files. Windows, a general operating system, tries to find the allocated space for a file from the file system so that it is not fragmented when saving a file, but in the following cases, a file may be saved as two or more non-adjacent files or divided files[12].

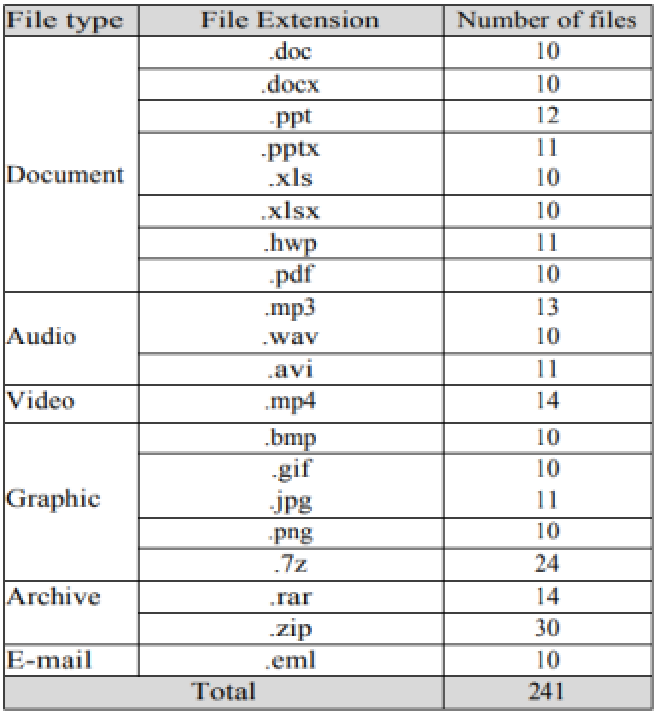
- When using the storage medium for a long period of time and adding or deleting multiple files when the capacity is almost full,
- When data is added from an existing saved file and there is insufficient sector space to store data at the end of the file,
- When the file system itself does not support writing files of a specific size in a continuous manner,
- When multiple processes take turns performing synchronous write operations.

As such, as the file system becomes outdated or files are created, modified, and deleted, fragmentation increases, and large-capacity files become more fragmented. Usually, people edit a downloaded file for document work from home and save it under a different name, or edit and overwrite the file being created several times and save it.

In the case of large files, most of them are high-resolution still images or audio/video files, and it is not uncommon for these files to become fragmented. For storage, in this case, the files may be naturally fragmented and stored in the storage medium. In addition, fragmented and saved files can be divided into two categories: linear or non-linear. Linear fragmentation is a case in which a file is divided into two or more pieces, but the pieces are in a data set in order, and non-linear fragmentation is a case in which a data set exists in a different order from the original file

- **File fragmentation test method**

In order to reflect the realistic environment, we did not use a tool that artificially fragments the file, and when the file is fragmented into two or more, it is a case where fragmentation is easily performed. The process of inserting data was selected. The types and number of previously saved sample files correspond to Table 3, and the files for each type are varied in size between 2KB and 55MB. In order to fragment the saved sample file, large data is inserted from the file and then the file is overwritten or saved with a different name. For document files except PDF, text, video, audio, graphic, and PDF files are inserted, and PDF files are overwritten with other PDF files through the merge function. In the case of audio files and video files, merge or reduce other audio and video files through an editing program, then overwrite and save as a different name. In the case of graphic files, the file is overwritten by increasing the size of the picture through an editing program and editing various effects. Finally, in the case of compressed files, video, audio, images, compressed files, and PDFs are inserted and the files are overwritten[13].



File type	File Extension	Number of files
Document	.doc	10
	.docx	10
	.ppt	12
	.pptx .xls	11 10
	.xlsx	10
	.hwp	11
	.pdf	10
Audio	.mp3 .wav	13 10
	.avi	11
Video	.mp4	14
Graphic	.bmp	10
	.gif .jpg	10 11
	.png	10
	.7z	24
Archive	.rar	14
	.zip	30
E-mail	.eml	10
Total		241

Fig 2. Sample File type for Carving Verification

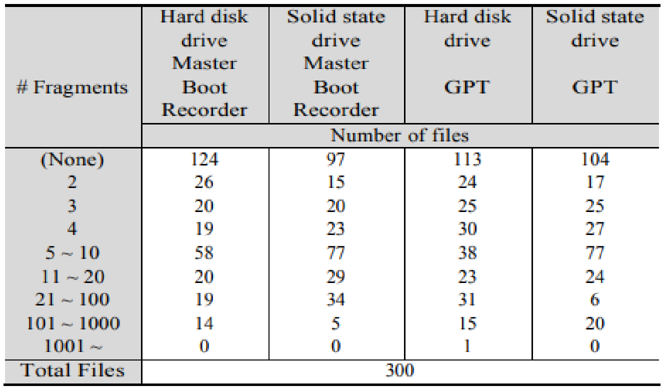
- **File Fragmentation Experiment Results**

As a result of completing the scenario for file fragmentation among a total of 241 files by the above experimental method in each configured PC environment, the number of fragmented files was confirmed through the file data run list. Since the file attribute content is larger than the size of the MFT entry, most non-resident attribute files are allocated non-contiguously when there is no free space, which is a method to receive and store a separate cluster.

In order to effectively manage these non-contiguously allocated clusters, it is called a ”cluster run,” and the expression of the cluster run is called a ”run list.” Figure 3 shows the number of fragmented files from each file. It can be seen that many files are fragmented into two or more pieces and stored, and the fragmentation of files is greater in SSD than in HDD[14].

3.2.2 Delete files

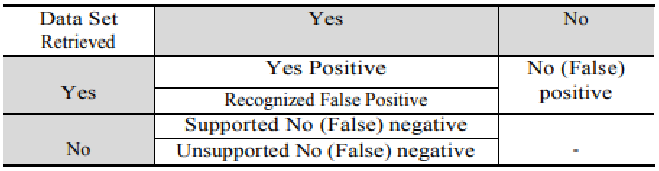
In most cases, files are deleted during the investigation of digital devices, and the recovery of deleted files plays an important role in forensics. Data deletion from storage media can be classified into software-based deletion, hardware based degaussing, and



# Fragments	Hard disk drive Master Boot Recorder	Solid state drive Master Boot Recorder	Hard disk drive GPT	Solid state drive GPT
	Number of files			
(None)	124	97	113	104
2	26	15	24	17
3	20	20	25	25
4	19	23	30	27
5 ~ 10	58	77	38	77
11 ~ 20	20	29	23	24
21 ~ 100	19	34	31	6
101 ~ 1000	14	5	15	20
1001 ~	0	0	1	0
Total Files	300			

Fig 3. Fragmentation results of files

physical-based destruction. Software-based erasure is a method developed for HDDs that typically writes a specific data pattern to each sector of the disk in a sequential manner, overwriting the original data and making it unrecoverable. However, since the SSD is controlled by the wear-leveling controller, the software cannot control the specific area where the data is recorded. As an alternative to software-based erasure, hardware-based erasure works by sending a magnetic pulse through the medium to the degauser. In most cases, this is a quick way to render an HDD inoperable, but in the case of an SSD, it is not an effective method because data is stored electronically rather than magnetically[15]. The best way to erase data from HDD and SSD drives is physically based destruction, which is usually the process of shredding a single chip into very small pieces. We deleted the data set file in the storage medium through the USB boot format, which is generally generated for data deletion during the digital investigation process.



Data Set Retrieved	Yes	No
Yes	Yes Positive	No (False) positive
	Recognized False Positive	
No	Supported No (False) negative	-
	Unsupported No (False) negative	

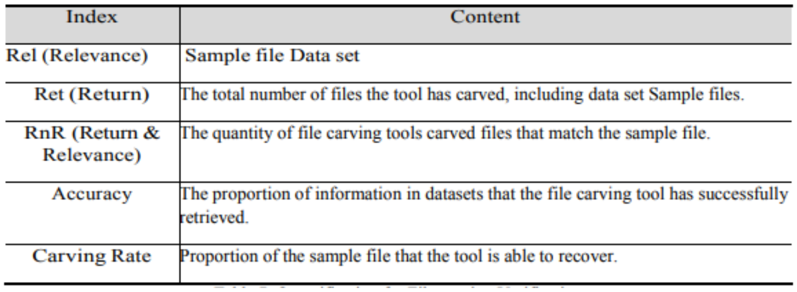
Fig 4. Quality of the carving results

3.3 File Carving Tool Verification/Result & Discussion

3.3.1 File Carving Results

In verifying the file carving tool, information on the layout of the sample file set was collected. This information consists of a list of all files, the size of the sample file, the block range, and the MD5. MD5 represents a 32-character hexadecimal number calculated using a cryptographic hash function, which can be used to uniquely identify a specific file. When carving files from tools, the results can lead to three main types. Figure 4 shows three types of results. If the carving result matches MD5 of the sample file set, it is judged to be a correctly carved file and means positive. However, MD5 does not mean false positive if there is no mismatch. Most file types have a certain amount of space at the beginning and end of the file. For example, an html file is carved with new empty data that does not need to be carved for correct results to generate different MD5 values. However, if the block range is the same, the carving result can be called positive, and this result is known as ”false positive.” Therefore, we use SSD deep, a file similarity measurement tool, to judge the case where the similarity between files is greater than 99% as the same as positive[16].

In addition, results other than positive and known false positives can be called false negatives. That is, a case included in the sample file set but not carved is indicated as a false negative. In addition, false negative can be divided into a case where the tool does not support the sample file, so it is not carved, and a case where the tool does not sup-port the sample file, but the carving process cannot be performed properly[17]. These are called ”supported false negatives” and ”unsupported false negatives,” respectively. The tool result is that supported false negatives have lower reliability than unsupported false negatives, so it should be judged as bad. In order to quantify the file carving results and determine the standard of functionality, the



Index	Content
Rel (Relevance)	Sample file Data set
Ret (Return)	The total number of files the tool has carved, including data set Sample files.
RnR (Return & Relevance)	The quantity of file carving tools carved files that match the sample file.
Accuracy	The proportion of information in datasets that the file carving tool has successfully retrieved.
Carving Rate	Proportion of the sample file that the tool is able to recover.

Fig 5. Quantification for File carving Verification

dataset (sample file) is ex-pressed as Rel (abbreviation for Relevance), and the number of files recovered from the tool (positive, known false positive) is expressed as Ret (abbreviation for Return) was expressed. And among the files recovered from the tool, the number of files corresponding to the sample files was set as RnR (Return & Relevance). To determine the performance of the tool from these numerical values, the accuracy of the tool is calculated as, indicating the ratio of data successfully carved by the tool from the dataset (sample file). The rate of carving is calculated as it represents the ratio of data sets (sample files) among the data carved by the tool. Figure 4 provides an overview of the digitization for file carving verification[18]. In the case of media created through the file fragmentation scenario, in addition to the sample files in Figure 2, the results are calculated by including the number of files saved under other names, files included in compressed files, and large files inserted or merged for fragmentation.

- **Carving Rate**

Figures 2 and 3 show the carving rate and accuracy of the Foremost tool for each scenario in Master Boot Recorder Hard Disk Drive, Master Boot Recorder Solid State Drive, GPT Hard Disk Drive, and GPT SSD storage media, respectively. It can be seen that the carving rate shows a comparatively higher result than the accuracy, but the overall result is quite low. We did not directly designate the file information of the target to be carved from the Foremost configuration file but used the existing supported method. Therefore, the reasons for the low carving results are as follows: MP4 and eml file signatures are not supported. The header signature is present but the footer signature is not for xlsx, pptx, docx, mp3, wav, avi, bmp, and png files. In this case, it is impossible for the tool to accurately recover files by specifying an arbitrary maximum file size. There are cases where the size of the sample file is larger than the randomly assigned maximum file size. In the scenario (S1) where the data set is neither fragmented nor has the disk been formatted, this is to measure the file carving performance in the storage medium where the file is linearly stored. The carving rate was 25.2%, showing the best results in GPT SSD storage media, and Master Boot Recorder Solid State Drive, GPT Hard Disk Drive, and Master Boot Recorder Hard Disk Drive show high performance in turn. In particular, it appears to be relatively better in SSD than in HDD and in GPT than in MBR. Overall, linearly, the carving performance from the storage medium where the file is stored was better than in the other scenarios, which means that the tool ignores the file system and reads the data sequentially from the first offset of the storage medium, starting with the header and processing of the file. It seems that this is because it is effective in recovering linearly stored files because it is a recovery technique using data. In the scenario (S2) in which the storage medium of S1 is formatted, the result is as low as 0.5% in the GPT HDD when measuring the file carving performance when the file is deleted. In addition, all three other storage media failed to carve the sample file at 0%. As mentioned in 4.1, when a file is deleted, the carving performance may be lower because the file may be fragmented in a specific area or data may be erased. When the storage medium is formatted like S2, it is difficult to carve files from Foremost, but it seems that the file carving performance of HDD is better than SSD when performing trim operation.

The scenario (S3) where the dataset is stored to measure, the disk is fragmented and not formatted the non-linearly stored file carving performance. The carving rate was 0% in the MBR structure, and the HDD and SSD in the GPT structure. It was the same as 1.9%. However, all files recovered from the storage medium are non-fragmented files, and the carving rate of the tools targeting the stored files is 0%, indicating that sample files cannot be recovered. Finally, when the data set is fragmented and the disk is formatted (S4), the file is fragmented, stored, and then deleted.

- **Accuracy**

Figure 3 shows the accuracy of the Foremost tool for each scenario in Master Boot Recorder, Hard Disk Drive, Master Boot Recorder Solid State Drive, GPT Hard Disk Drive, and GPT SSD storage media, respectively. Overall, the accuracy is very low,

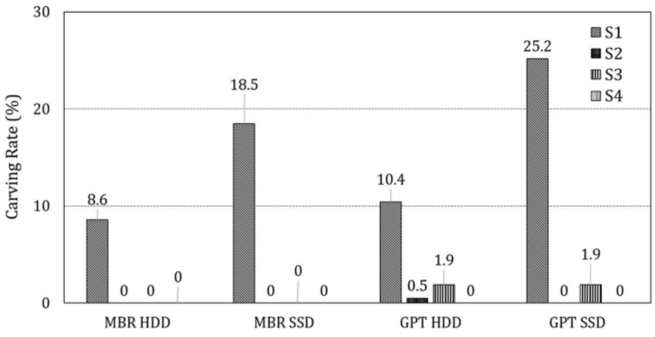


Fig 6. File Carving Ratio of Foremost Tool (Range Zoomed: 30%)

with a result of less than 10%. First, in the scenario in which the dataset is not fragmented the disk has not been formatted, (S1), that is, when the files on the storage medium are stored linearly, the GPT SSD was the best at 0.52%, although the result was not high, and the overall result was the same as the carving rate. As a result, the carving performance from the storage medium where the files are linearly stored was better overall than in the other scenarios. It appears to be relatively better in SSD than in HDD and in GPT than in MBR.

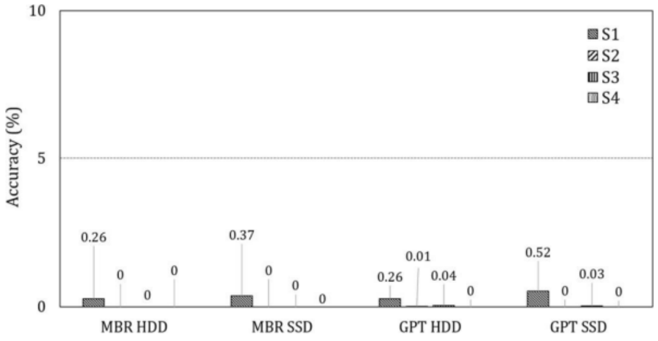


Fig 7. Accuracy ratio of foremost tool (Range Zoomed: 10%)

The scenario (S2) in which the storage medium of S1 is formatted is to assess the effectiveness of the file carving when a file is deleted, and while the accuracy of most of the storage medium carving fails, one compressed file was carved with 0.01% in the GPT HDD. It may indicate that the HDD has a higher probability that files can be carved than on the SSD. The scenario (S3) in which the dataset is stored unformatted, and the drive is fragmented is to measure the non-linearly stored file carving performance.

The accuracy was 0% in the MBR structure, and the accuracy in the GPT structure was 0.04, respectively. % and 0.03%, which shows very low results. However, as in the carving rate, all the files recovered from the storage medium are non-fragmented files, and it can be seen that the accuracy of the tools targeting the stored files is 0% and cannot be recovered. Finally, when the data set is fragmented and the disk is formatted (S4), the file is fragmented, stored, and then deleted as a storage medium. The accuracy of all storage media is 0%, and they failed to carve the sample file. As a result of the experiment, the performance of Foremost is quite low, and since it is a carving tool that recovers using the header and footer of the file, the result is relatively good in the scenario (S1) where the file is not fragmented and the file is saved. Failed to carve fragmented files in fragmented scenarios (S3, S4).

In reality, important files in digital forensics are easily fragmented and stored on the storage medium, and in the case of a person with a malicious purpose, most of the actions are to format the storage medium and hide it. It was found in the tool verification that the carving results in the storage medium reflecting these various scenarios were quite poor, and various tools perform file carving based on Foremost or Scalpel. When you perform a recovery, you may find that it can be difficult to help you solve the problem. In the same scenario, the performance of Foremost's file carving tools is different in HDD and SSD, and MBR and GPT, and most of the carved files have relatively many documents and images, so it is more helpful in recovering images and document files. It seems that this could be also, since there are relatively more files that match more than 99%

through ssd deep than when the hash values match, it can be seen that it may be difficult to recover exactly the same file.

4 Conclusion

In this study, the limitations of the dataset developed from the existing research and projects for file carving tool verification were pointed out, and images reflecting realistic environments and scenarios were developed. The environment of the developed image consists of four storage media environments consisting of MBR, GPT, HDD, and SSD; disk media in which files are not fragmented; disk media formatted after files are stored without fragmentation; and disk media in which files are stored fragmented. The files were fragmented and saved, and then the formatted disk media created images according to a total of four scenarios. As a result, foremost v1.5.7, a well-known file carving tool, was used to quantify the results for the verification of 16 images developed. Although there is a limitation that it is limited to only the foremost tool, in future research, it is possible to perform the carving tool verification effectively by further subdividing the scenario image and to verify the various carving techniques and tools studied using the developed image. By comparing them, it is expected to contribute to the research and development of better performing carving techniques and tools by comparing them.

In conclusion, our investigation has provided a comprehensive exploration of file carving tool methodologies and the pivotal role of scenario-based image creation in digital forensics. The findings and insights gained from this research contribute to the ongoing evolution of digital forensic techniques and tools, ultimately enhancing the field's capabilities in data recovery, investigation, and cybersecurity. As digital technologies continue to advance, the digital forensics community must remain agile and adaptive, seeking innovative solutions to meet the challenges of the digital age.

5 Abbreviation

SSD — Solid State Drive; HDD — Hard Disk Drive; CFTT — Computer Forensics Tool Testing; NIST — National Institute of Standards and Technology; NTFS — New Technology File System; CFReDS — Computer Forensic Reference Data Sets; BIOS — Basic Input/output System; GUID — Globally Unique Identification; GPT — GUID Partition Table; EFI — Extensible Firmware Interface; RDC — Real Data Corpus; MBR — Master Boot Record

References

1) Alherbawi N, Shukur Z, Sulaiman R. Systematic Literature Review on Data Carving in Digital Forensic. *Procedia Technology*. 2013;11:86–92. Available from: https://doi.org/10.1016/j.protcy.2013.12.165.
2) Forensic Images for File Carving. . Available from: https://cfreds-archive.nist.gov/FileCarving/TestFiles/index.html.
3) Digital (Computer) Forensics Tool Testing Images. . Available from: https://dftt.sourceforge.net/.
4) Disk images - digital corpora. . Available from: https://digitalcorpora.org/corpora/disk-images/.
5) Casey E. Digital Evidence and Computer Crime: Forensic Science, Computers and the Internet. 3rd ed. Academic Press. 2019. Available from: https://booksite.elsevier.com/samplechapters/9780123742681/Front_Matter.pdf.
6) Basic Data Carving Test #1. . Available from: https://dftt.sourceforge.net/test11/index.html.
7) Hadi HJ, Musthaq N, Khan IU. SSD Forensic: Evidence Generation and Forensic Research on Solid State Drives Using Trim Analysis. In: 2021 International Conference on Cyber Warfare and Security (ICCWS). IEEE. 2022;p. 51–56. Available from: https://doi.org/10.1109/ICCWS53234.2021.9702989.
8) Kim H, Kim J, Kwon T. A study of verification methods for File Carving tools by scenario-based image creation. *Journal of the Korea Institute of In-formation Security and Cryptology*. 2019;29(4):835–845. Available from: https://doi.org/10.13089/JKIISC.2019.29.4.835.
9) Suthar H, Sharma P. An Approach to Data Recovery from Solid State Drive: Cyber Forensics. In: Advancements in Cybercrime Investigation and Digital Forensics. Apple Academic Press. 2023;p. 1–20. Available from: https://www.appleacademicpress.com/advancements-in-cybercrimeinvestigation-and-digital-forensics-/1119.
10) Et MH. Comparative Analysis Study on SSD, HDD, and SSHD. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021;12(3):3635–3641. Available from: https://turcomat.org/index.php/turkbilmat/article/view/1644.
11) Suthar H, Sharma P. Guaranteed Data Destruction Strategies and Drive Sanitization: SSD. *Research Square*. 2022;p. 1–15. Available from: https://assets.researchsquare.com/files/rs-1896935/v1/75292483-17ad-40b7-ad59-9b14557e6236.pdf?c=1660541356.
12) Suthar H, Sharma P. Method for Extracting Data from an Overprovisioned SSD. In: 2022 IEEE Pune Section International Conference (PuneCon). IEEE. 2023;p. 1–6. Available from: https://doi.org/10.1109/PuneCon55413.2022.10014904.
13) Afifi M. State-of-the-Art Tools and Techniques in Digital Forensics. In: Advances in Digital Forensics XV . Springer. 2019;p. 29–43. Available from: https://www.springeropen.com/collections/advdigitalforensics.
14) Javed AR, Ahmed W, Alazab M, Jalil Z, Kifayat K, Gadekallu TR. A Comprehensive Survey on Computer Forensics: State-of-the-Art, Tools, Techniques, Challenges, and Future Directions. *IEEE Access*. 2022;10:11065–11089. Available from: https://doi.org/10.1109/ACCESS.2022.3142508.
15) Ali RR, Mohamad KM. RX_myKarve carving framework for reassembling complex fragmentations of JPEG images. *Journal of King Saud University - Computer and Information Sciences*. 2021;33(1):21–32. Available from: https://doi.org/10.1016/j.jksuci.2018.12.007.
16) Suthar H, Sharma P. SSD Forensic Investigation Using Open Source Tool. In: Examining Multimedia Forensics and Content Integrity. IGI Global. 2023;p. 56–78. Available from: https://www.igi-global.com/chapter/ssd-forensic-investigation-using-open-source-tool/324147.
17) Suthar H, Sharma P. A Technique for decreasing the SSD's Garbage Collection overhead using ML techniques. *International Conference on Science, Engineering and Technology (ICSET 2022)*. 2023;p. 51–58. Available from: https://soe.rku.ac.in/conferences/data/06_9738_ICSET%202022.pdf.

18) Oh J, Lee S, Hwang H. Forensic Recovery of File System Metadata for Digital Forensic Investigation. *IEEE Access*. 2022;10:111591–111606. Available from: https://doi.org/10.1109/ACCESS.2022.3213030.