ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

UDC 004.63.056.3 **doi: 10.32620/reks.2023.3.16**

Maksym BOIKO[1,2], Viacheslav MOSKALENKO[1,3], Oksana SHOVKOPLIAS[1]

[1] ***Sumy State University, Sumy, Ukraine***
[2] ***The National Anti-Corruption Bureau of Ukraine, Kyiv, Ukraine***
[3] ***National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine***

ADVANCED FILE CARVING: ONTOLOGY, MODELS AND METHODS

File carving techniques are important in the field of digital forensics. At the same time, the rapid growth in the amount and types of data requires the development of file carving methods in terms of capabilities, accuracy, and computational efficiency. However, most of the methods are developed to solve specific tasks and are based on a certain set of assumptions and a priori knowledge about the files to be recovered. There is a lack of research that systematizes methods and structures approaches to identify gaps and determine perspective directions for development, considering the latest advances in information technology and artificial intelligence. The ***subject matter*** *of this article is the structure, factors, efficiency criteria, methods, and tools of file carving, as well as the current state and tendencies of development of file carving methods. The* ***goal*** *of this study is to systematize knowledge about advanced file carving methods and identify perspective directions for their development. The* ***tasks*** *to be solved are as follows: to identify the main stages of file carving and analyze approaches to their implementation; to build an ontological scheme of file carving; and to identify perspective directions for the development of carving methods. The* ***methods*** *used were literature review, systematization, and summarization. The obtained* ***results*** *are as follows. An ontological scheme for the file carving concept is constructed. The scheme includes the principles, properties, phases, techniques, evaluation criteria, tools used, and factors influencing file carving. The features, limitations, and fields of application of the data recovery methods are provided. It was established that the most widespread approach to file reconstruction is still a manually detailed analysis of the internal structure of files and/or their contents, identifying specific patterns that allow reassembling the sequence of data fragments in the correct order. However, most of the methods do not provide one hundred percent guaranteed results. This article analyzes the current state and prospects of using artificial intelligence methods in the field of digital forensics, particularly for identifying data blocks, clustering, and reconstructing files, as well as restoring the contents of media files with damaged or lost headers. The necessity of having priori information about the file structure or content for successfully carving fragmented data is determined.* ***Conclusions.*** *The scientific novelty of the obtained results is as follows: for the first time, advanced file carving methods are systematized and analyzed by directions of development and the perspectives of using artificial intelligence for identifying data blocks, clustering, and file content restoration; for the first time, an ontological scheme of file carving is constructed, which can be used as a roadmap for developing new advanced systems in the digital forensics field.*

Keywords: *digital forensics; metadata; fragmentation; fragmented file; data recovery; file carving; file fragment identification; file reconstruction; file restoring; artificial intelligence.*

1. Introduction

1.1 Motivation of research

Users constantly create, view, edit, and delete many files when working with data. This is a dynamic process. The file system is responsible for the mechanisms and rules for storing data on the disk space [1]. Researchers regularly search for deleted information and recover it when conducting digital forensic examinations. This is explained by the fact that when illegal or compromising activities are performed, it is evident that users try to cover their trails and delete any sensitive information.

If it does not consider the SSD's internal processes [2], file systems usually optimize their work so that they do not take any action with deleted data blocks [1]. Such disk space areas are only marked as free for use and remain intact until they are allocated for storing other information. As a result, unallocated disk space can contain forensically important data.

Some file types (for example, TXT, LOG, DOC) store their data in an uncompressed form. Their full or partial contents can be accessed without restoring the entire object by reading their detected data blocks or identifying text fragments using search terms. However, this is not sufficient when trying to extract the contents of compound files that use compression, encryption, or have a complex internal structure. These file types include JPG, BMP, AVI, MPG, DOCX, XLSX, PDF, and SQLITE.

A separate digital forensics sphere is the study of RAM, particularly volatile memory dumps in the

Windows operating system [3]. RAM areas may contain the contents of files the user has been working with that may not have been stored on the disk [4]. Such files occupy non-contiguous data blocks, the location of which may not always be known.

The file recovery process is more difficult when the files are fragmented and there is no file allocation data. In the above circumstances, searching for file fragments and their corresponding positioning is a time-consuming and complex task with unclear solutions. In addition, it is necessary to consider the increasing number of digital devices and the amount of information available in general. In recent years, there has been an intensifying use of advanced file carving techniques to solve and optimize various stages of such tasks.

1.2. Research gap

In recent years, researchers have periodically reviewed file carving techniques. The most common methods of data recovery are presented in [5].

Some authors have focused on a survey of various data carving techniques of multimedia [6, 7] or JPEG [8] files. In [9], the researchers focused only on the efficiency analysis of Scalpel and Foremost carving processes. The paper [10] discusses the recovery of a more extensive set of file types focused on fragmented Microsoft Word documents.

In other cases, the file carving algorithms were divided according to a particular principle. For example, in [11], the authors classified carving methods for JPEG files into basic and advanced categories and conducted a detailed analysis of graph theoretic and weightage techniques. Similar approaches to the classification of file carving methods are used in [12], where the author also presents a taxonomy of file carving techniques.

In addition to the techniques and carving directions discussed, the work [13] includes data recovery research area mapping.

Despite a relatively large number of surveys on data recovery techniques, the authors did not comprehensively consider the problem of file carving. The works are not sufficiently systematized. In addition, the ontological relationship between file carving and various aspects of this process has not been established.

1.3. Objectives and Contributions

This study systematizes and build schemes for recovering highly fragmented files using advanced techniques and determine the feasibility of using artificial intelligence in this process.

The key issues are as follows:

- analyze the existing advanced file carving techniques;
- identify the stages of data recovery where these techniques are applied;
- determine the feasibility of using artificial intelligence and advanced techniques.

Structurally, this work consists of the following sections. The research methodology is described in section 2. Section 3 discusses the main phases of digital forensics, data recovery with and without file system metadata, and the ontological diagram of file carving. Advanced file carving techniques and their details are provided in section 4. Section 5 presents a discussion of the aforementioned techniques. The last section provides conclusions and indicates directions for future research.

2. Research methodology

The research hypothesis is that the carving of highly fragmented files depends on three key factors:

- from improving the efficiency of the identification of data fragments in unallocated space and/or RAM;
- from the techniques of reconstruction of the detected file fragments;
- directly from the file type and its internal content.

For this purpose, the three research questions identified for the current literature review are shown in Table 1.

Table 1

Research questions

#	**The question**
Q1	What are the typical stages of file carving, and what are the perspectives for improving each stage?
Q2	Is it possible to carve fragmented files without a priori information about their internal structure and contents?
Q3	What are the perspectives on using artificial intelligence methods in the file carving field?

The search approach is based on selecting and analyzing articles that address the problems of carving highly fragmented files or solve individual phases of this process.

The selection process consisted of several stages. Initially, the most relevant studies were identified by searching for keywords in the titles and abstracts. The articles were then reviewed for their relevance to the research questions. The final set of articles was based on the quality of the content.

For a complete understanding of the problems that appear when recovering fragmented data in the absence of file system metadata, the literature for the period of dynamic digital forensics growth approximately the last 20 years was analyzed.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

3. Background, Directions and Ontology

3.1 Digital forensics

Conducting digital forensics examinations, researchers perform several actions depending on the type of research, type and number of objects, tasks to be solved, etc. [14]. In general, this process can be conditionally divided into four stages: collection, examination, analysis, and presentation (Fig. 1).

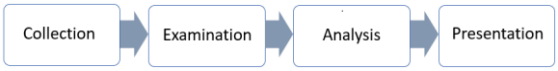

Fig. 1. Digital forensics stages

During the first stage, copies of digital media are collected and created. In the next phase, the created images are processed. As a rule, a full-fledged study of disk space is conducted: file system analysis, hidden information detection, deleted file recovery, signature analysis, indexing, pattern search, etc. In the last two stages, investigators identify important data, interpret them, and generate a report with detailed answers to the questions.

3.2 Deleted file recovery

The complexity of the deleted data recovery process depends on the file system, the character of the user's actions when deleting information, the character and duration of further actions, etc. (Fig. 2).

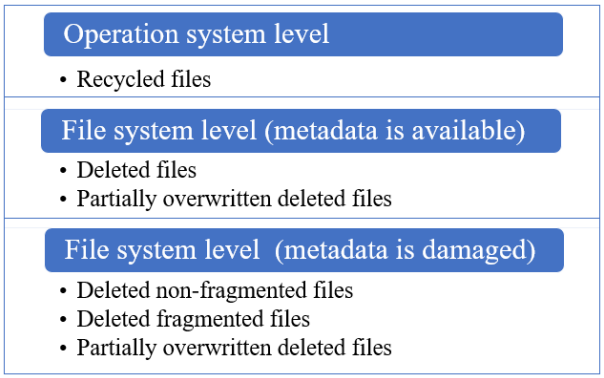

Fig 2. Data recovery by the degree of complexity (from the easiest at the top to the most difficult at the bottom)

The simplest case for recovering files in the most popular file systems (NTFS, FAT32, EXFAT, HFS, EXT) is to delete data from the Recycle Bin. In this case, the file is not actually deleted but is moved to another location. Therefore, the blocks of data it occupies and its metadata remain intact.

If the user deleted a file bypassing the Recycle Bin or emptied the latter one, two situations are possible:

- the file system metadata is not affected;
- metadata of the deleted file is lost.

If the metadata is available, the file can be recovered using information about the location of its data blocks [1]. The only nuance may be overwriting certain areas of the deleted file with other data. Then, at best, only a partial reconstruction of the file is possible with the subsequent loss of some or all of its contents, depending on the file type, number, and character of lost fragments.

Figure 3 illustrates a possible case of data overwriting. At the top is the initial state of the disk space with existing files #1, #2, and #3. At the bottom is the current state of the exact locations of the disk space, where file #1 is wholly overwritten and file #3 is partially overwritten after user manipulations. In this case, if the file system metadata is available, files #2 and #4 will be fully recovered, file #1 will be lost, and file #3 will be restored but partially overwritten. At the same time, the recovery of even partial contents of file #3 is highly questionable.

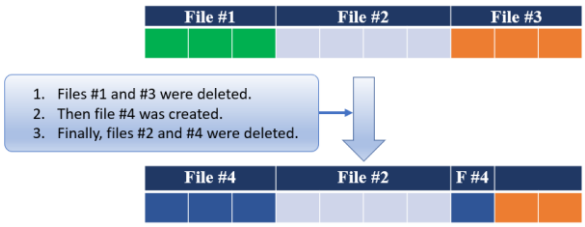

Fig. 3. Example of possible data overwriting

3.3 File carving

The biggest problems arise when recovering deleted information from lost or damaged file system metadata and information about the location of file data blocks. In this case, file carving techniques are applied, which are well suited for recovering contiguous files that contain a header and footer [15, 16]. For these purposes, the unallocated space is searched for file beginning and end signatures. However, this method has disadvantages if the file consists of two or more non-contiguous fragments.

Fig. 4 shows an imaginary example of locating data blocks of two deleted fragmented files on the disk space. File A is divided into four clusters that are located out of order. The file B occupies three clusters and is divided into two fragments. During the recovery process, the most likely problem is identifying the A3, A4, B2, and B3 fragments. If A and B belong to the same file type, it is necessary to define the boundaries of each file. Finally, to recover the file A, it is necessary to arrange the fragments correctly.

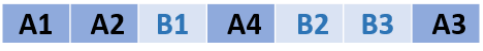

Fig. 4. Example of possible data fragmentation

Usually, it is not a problem to identify the first fragment of a file, which in most cases has a clear marker in the form of a header in its initial bytes. However, not all files have a footer, which can also have any offset relative to the beginning of the block/cluster. If the file consists of three or more fragments, the first key problem is to determine the data blocks that do not have clear markers, such as the header and footer. Subsequently, it is necessary to cluster the detected fragments and directly reconstruct the file or its contents.

Fig. 5 shows the ontological diagram, which indicates the principles of file carving, properties, tools required for this, the phases of file carving, factors that affect the result, techniques used, and criteria used for evaluation.

The set of software tools is shown in Fig. 5 contains only basic information, is not complete, and depends on the platform on which the data recovery process is performed. Usually, at the pre-processing stage of file carving, utilities such as FTK Imager, DD, X-Ways Forensics, and EnCase Imager are used to create a full bit-for-bit copy of the original media. Then, at the examination stage, the disk space is analyzed. For this purpose, universal tools such as X-Ways Forensics, UFS Explorer, EnCase, Magnet Axiom, Autopsy, Forensic Explorer, and FTK are most often used. They operate on the principle of a Swiss Army knife. Scalpel, Foremost, PhotoRec, and RecoverIt are utilities explicitly designed for data recovery, which is performed using proprietary algorithms. The abovementioned software does not guarantee 100% results and works well only with non-fragmented data. For this reason, in non-trivial cases, specialists often use additional tools for manual data recovery, such as Hex Editors and highly specific scripts [17].

To evaluate the effectiveness of the software, it is advisable to determine the number of correctly recovered (true positive or TP), incorrectly recovered (false positive or FP), and unrecovered (false negative or FN) files [18]. Subsequently, the following criteria are applied:

- precision – the percentage of correctly recovered files among the results of the utility's work [18]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}; \qquad (1)$$

- recall – the percentage of correctly recovered files from their total number in the digital media [18]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}; \qquad (2)$$

- f-measure – the overall performance of a tool [18 - 20]:

$$\text{Fmeasure} = \frac{1}{\alpha/\text{P}+(1-\alpha)/\text{R}}, \qquad (3)$$

where P is the precision, R is the recall, α is the numeric value from 0 to 1 used to determine precision and recall weights;

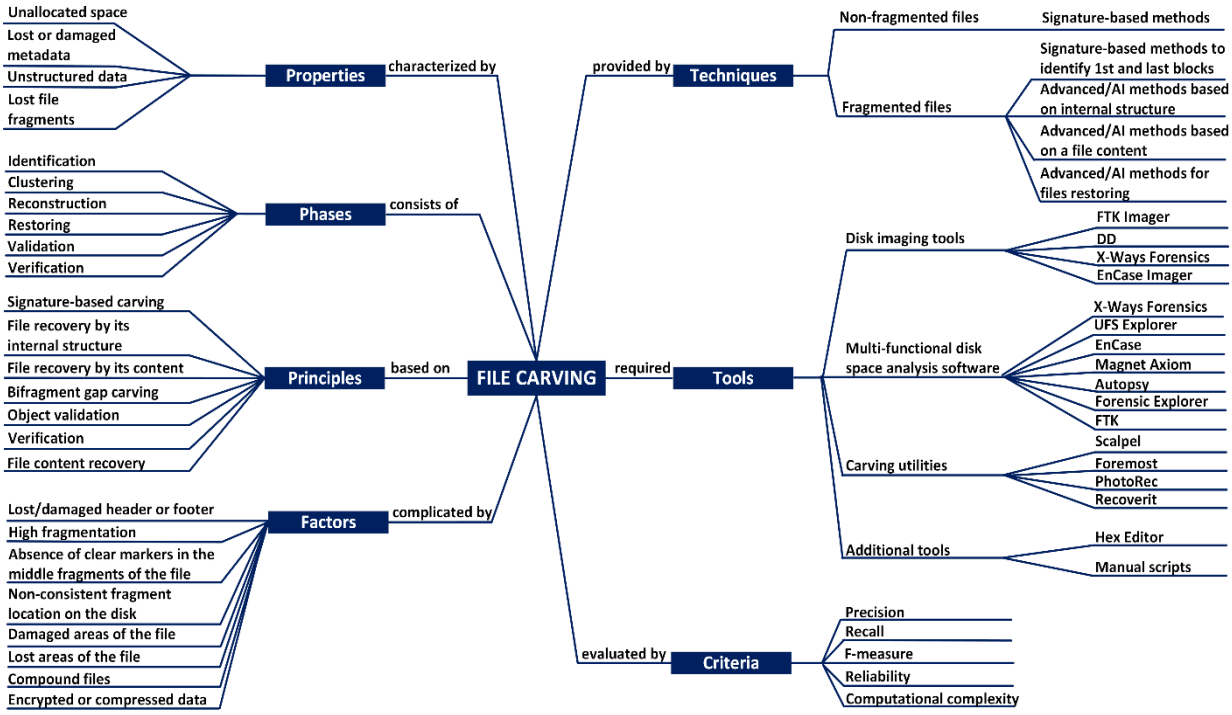

Fig. 5. The ontological diagram of file carving

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

- reliability – the tool's efficiency among supported file types [18 - 20]:

$$\text{Reliability} = \frac{\text{SF}-\text{SFN}}{\text{SF}}, \quad (4)$$

where SF is the number of supported files in the dataset and SFN is the number of supported false negatives;

- computational complexity is the amount of resources required to solve the task. Computational complexity is often estimated by the data processing speed or task execution time with the same computing resources [19, 20].

When comparing the effectiveness of the utilities, some researchers [19, 20] also divided false positive files into two categories: partially recovered (known false positive or kFP) and remaining files (unknown false positive or uFP). As a result, precision and recall are defined as follows [19, 20]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{uFP}+\text{kFP}/\beta}, \quad (5)$$

where β is the numeric value not less than 1 used to determine the relative weight of uFP compared with kFP;

$$\text{Recall} = \frac{\text{all}-\text{FN}}{\text{all}}, \quad (6)$$

where all is the total number of files in the dataset.

Each of the above metrics (Precision, Recall, F-measure, Reliability) can take a value from 0 to 1 and show the quality of the tool. Metric values close to 1 indicate that the software shows good performance. Table 2 shows the interpretation of the low values of Precision, Recall, and Reliability metrics [18 - 20]. It is worth noting, the authors often compare the number of successfully recovered files using their methods with the results of recognized utilities such as Scalpel, Foremost, PhotoRec, etc.

Table 2

Interpretation of the low values of the metrics

Metric	Interpretation
Precision	A large number of false positives
Recall	A small number of correctly recovered files
Reliability	A large number of fails when recovering supported file type

4. Advanced file carving techniques

To review advanced file carving techniques, we analyzed the works available on resources such as ScienceDirect, Elsevier, and IEEE. To do this, a search was conducted using the keywords and their combinations shown in Table 3.

Table 3

Search terms

#	Keywords
1	file carving
2	data carving
3	smart carving
4	machine learning
5	artificial intelligence
6	data recovery
7	fragmented files

The most relevant detected works, their direction, brief description, and particularities are shown in Table 4. In general, from these studies, advanced file carving techniques are successfully used to varying degrees at the identification, clustering, reconstruction, and restoration stages in addition to standard digital forensics methods. However, most studies do not clearly distinguish between these phases. For example, clustering and validation often occur during file reconstruction and/or restoring. Usually, these issues are solved in parallel. In addition, most of the authors who addressed the issue of reconstruction or restoring performed data validation and verification. Therefore, the last two stages are not mentioned separately in Table 4.

In this case, identification means identifying data blocks related to a specific type of data or files. Clustering involves dividing the identified data fragments into groups of blocks belonging to different files. The identified data blocks are placed in the correct order during reconstruction. Instead, during the restoring process, the file's contents are restored in case of damage or loss of some file areas.

Fig. 6 shows pre-processing and typical stages of file carving.

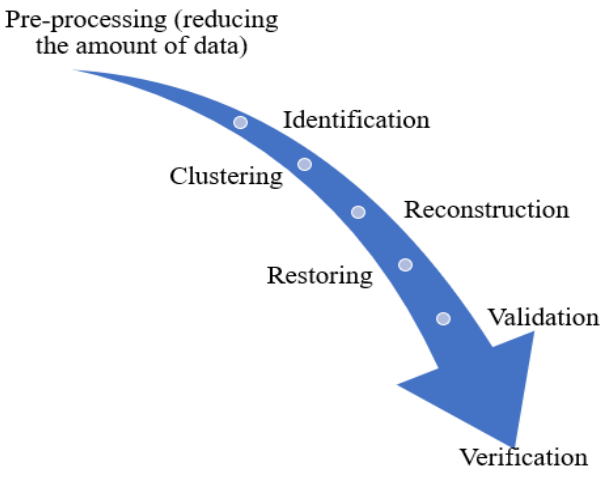

Fig. 6. Steps of pre-processing and file carving

Table 4

Advanced file carving techniques

Authors	Direction	Summary
Zanero [21]	Identification	Applying a set of support vector machines classifiers to determine data blocks for the files of the following types: BMP, DOC, EXE, GIF, JPG, MP3, ODT, PDF, PPT (9 classes). Average true positive rate – 90.4%, average false positive rate – 12.4%.
Fitzgerald et al. [22]	Identification	File fragment classification using a supervised learning approach based on support vector machines combined with the bag-of-words model (24 classes). The best results were obtained for CSV, PS, GIF, SQL, HTML, JAVA, XML, and BMP files (>90%). Fragments of PPTX, PPS, DOCX, XLSX, PPT, SWF, JPG, ZIP, GZ, PDF, and TXT files – 2.3% to 31.8% of prediction accuracy.
Beebe et al. [23]	Identification	Using support vector machines (N-gram vectors) to classify data blocks across 30 file types and 8 data types. Overall classification rate – 73.4%. High misclassification rate of encrypt, PPT, ZIP, PPTX, GZIP, PNG, FLV, DOC, XLSX, PDF, DOCX, AVI, and BMP files.
Pan et al. [24]	Identification	A method to identify the AVI-type blocks based on their internal structure. False positive rate – 53% (2 classes).
Wang et al. [25]	Identification	File fragment classification (18 classes) using N-grams frequencies. The average prediction accuracy is up to approximately 61%. Problems with classifying XLSX, PPTX, DOCX, GZ, PNG, PDF, PPT, and SWF files.
Karampidis et al. [26]	Identification	Comparison of machine learning methods (Decision Trees, Support Vector Machines, Neural Networks, Logistic Regression, k-Nearest Neighbor) for data block identification. Prediction accuracy – 89% to 100%. Only 4 different classes (JPG, PDF, PNG, GIF).
Al-Sadi et al. [27]	Reconstruction	Reconstructing graphic files by determining the image to which a fragment belongs. NaiveBayesMultinomialUpdateable, MultiClass, RandomForest, and BayesNet classifiers are used to determine the similarity between pixel values. The best results are 91% to 99.2% on average. Only graphic files.
Bhatt et al. [28]	Identification	File fragment classification using a hierarchical machine-learning-based approach with optimized support vector machines (SVM) 14 classes – CSV, DOC, HTML, PDF, PPT, XML, XLS, TXT, GIF, JPG, PNG, PS, SWF, and GZ. An average accuracy of 67.78%. PPT, PDF, DOC fragments – the worst results.
Sportiello et al. [29]	Identification	Construct SVM classifiers to determine the type of data block. 8 classes – BMP, DOC, EXE, GIF, JPG, MP3, ODT, and PDF files.
Mittal et al. [30]	Identification	512-byte and 4096-byte fragment type classification using convolutional neural networks with automatic feature extraction. 65.6% and 77.5% accuracy in the case of 75 classes. HEIC, MOV, 7Z, DMG, ZIP, EXE, PPTX, DJVU, PDF, DOCX – quite low rates.
Sester et al. [31]	Identification	File type identification approaches using support vector machines and neural networks for n-gram analysis. 6 classes – CSV, DOC, JPG, PPT, TXT, and XLS. Approximately 73% to 98% accuracy in different cases.
Chen et al. [32]	Identification	4096-byte fragment type classification using a deep convolution neural network. 16 classes – CSV, DOC, DOCX, GIF, GZ, HTML, JAVA, JPG, LOG, PDF, PNG, PPT, RTF, TEXT, XLS, and XML. 70.9% accuracy. Low results – DOC, DOCX, GIF, JPG, PNG, and TEXT. Represent all bytes of the data block as a grayscale image (automatic feature extraction).
Hiester [33]	Identification	Using recurrent (RNN), convolutional (CNN), and feed-forward neural networks (FNN) as classifiers of 512-byte data blocks 4 classes: CSV, XML, JPG, and GIF. Up to 98% accuracy in the best case (automatic feature extraction).
Ghaleb et al. [34]	Identification	512-byte and 4096-byte fragment type classification using light-weight convolutional neural networks. 66.33% and 79.27% accuracy in the case of 75 classes.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Continuation of Table 1

Authors	Direction	Summary
Liu et al. [35]	Identification	A 512-byte fragment type classification technique that converts the byte stream in a 2-D grayscale image and then captures both sequences by convolutional neural networks. 71.4% accuracy in the case of 75 classes.
Bharadwaj [36]	Identification	Using grayscale image conversion and convolutional neural networks to detect the compression algorithm of 4096-byte data block. 8 classes – rar, gzip, zip, 7-zip, bzip2, ncompress, lz4, and brotli. The achived accuracy is 41 % after five epochs.
Hague et al. [37]	Identification	Using the feature generation model, Byte2Vec, for feature extraction from 4096-byte fragments and k Nearest Neighbors for classification. 35 to 42 classes. An accuracy rate of 74%.
Vulinovic et al. [38]	Identification	File type identification using feed-forward and convolutional neural networks. 18 classes – CSV, DOC, DOCX, GIF, GZ, HTML, JPG, PDF, PNG, PPT, PPTX, PS, RTF, SWF, TXT, XLS, XLSX, and XML. Macro-average F1-score: FFNN – 79,93% to 81,38%, CNN – 61,55%.
Heo et al. [39]	Identification Restoring	Identification and restoration of damaged audio files using feed-forward and Long Short Term Memory (LSTM) neural network. High rates of identification of audio files.
Na et al. [40]	Identification Restoring	Restoring fragmented and partially overwritten video files by video frame analyses. 40 to 50% of the video with damaged data (50% overwriting) was recovered. Only MPEC-4 and H.264 video formats.
Amrouche et al. [41]	Restoring	Recover damaged images with a lost header. 90% accuracy of image properties identification; 78% accuracy for header prediction.
Alghafli et al. [42]	Identification Restoring	Identification and recovery of video with lost video codecs specifications. Problems with fragmented files.
Qiu et al. [43]	Identification Reconstruction	Using the byte frequency distribution and rate of change as features for building a classifier based on SVM. Reassembling fragments of the same file type using the PUP approach. The target file type is JPEG. Other file types are PNG, XML, HTML, PDF, GZ, ZIP, Office, MP3, and TXT. Better results (40.9% to 85.7%) compared with PhotoRec.
Guo et al. [44]	Identification Reconstruction	Using SVM for high-entropy file fragment classification and Parallel Unique Path algorithm for multimedia file reconstruction. Only 3 types (DOC, JPEG, C++ source code) were studied.
Ali et al. [45]	Identification Reconstruction	JPEG carving framework using an extreme learning machine and evolutionary algorithms for data block identification, validation, and reassembling. 90 to 93% accuracy. Problems with more than 2 fragmentation patterns or intertwined images.
Ali et al. [4]	Identification Clustering Reconstruction	Analysis of the textual contents of DOCX files in RAM and application of K-mean and Hierarchical clustering techniques to recover documents' texts. 54.35% to 90.54% of recovered documents. Possible problems with fragmented data blocks.
Al-Sharif et al. [46]	Identification Clustering Restoring	Finding PDF fragments in RAM using their internal structure. K-Means and Hierarchical clustering to define different documents. 46.34% to 50.24% of the PDF contents were carved (without file reconstruction).
Zhang et al. [47]	Identification Reconstruction	Finding and reassembling SQLite databases using knowledge of their internal structure. Time-consuming method.
Hilgert et al. [48]	Identification Reconstruction	Finding and reassembling PNG files using knowledge of their internal structure. Better results compared with PhotoRec, Scalpel, and Foremost. Problems with recovering files with missing fragments in the middle and/or the peculiarities of dividing the file into data blocks.
Tang et al. [49]	Reconstruction	Carving of highly fragmented JPEG files. The proposed framework can recover 97% of fragmented JPEG files. Fragmentation points are detected using the coherence of Euclidean Distance.

Continuation of Table 1

Authors	Direction	Summary
Ravi et al. [50]	Reconstruction	Carving fragmented text and some graphic files. Only several graphic file types (JPG, PNG, GIF). TXT files – dictionary-based approach.
Roussev et al. [51]	Identification	Presenting several file fragmentation techniques. The need to manually examine files and find specific features.
Lin et al. [52]	Reconstruction	DOC files' carving method based on internal structure. Better results (95,45%) than PhotoRec, Foremost.
Birmingham et al. [53]	Reconstruction	Carving fragmented JPEG files using knowledge about their internal structure. Better results compared with Adroit, FTK 3.3, Scalpel, PhotoRec, ProDiscover, and Encase 6. Does not cover out-of-order fragmentation.
Durmus et al. [54]	Reconstruction Restoring	Reassembling orphaned JPEG fragments using PRNU fingerprints of the cameras. It can also partially collect photos. 42% to 57% fragment localization accuracy
Chang et al. [55]	Reconstruction	JPEG fragment carving using pixel similarity. Success rate – 92%.
Uzun et al. [56]	Restoring	An Advanced Carver for JPEG Files. Ability to recover JPEG files with damaged or lost headers.
Boiko et al. [57]	Reconstruction	Reconstructing highly fragmented OOXML files. Up to 83% recovered files. Problems with embedding in documents.
Hand et al. [58]	Reconstruction	Utility for recovering binary executable files using their internal structure.
Xu et al. [59]	Identification Reconstruction	Identification and reassembly of EVTX Log fragments using their internal structure.
Garfinkel [16]	Reconstruction	Fast object validation for bi-fragmented files (JPEG, DOC, and ZIP files).

5. Discussion

As seen in Table 4, researchers have been quite successful in applying advanced methods to improve the mechanisms of deleted data recovery. The authors pay the most attention to the problem of fragment type identification, the general principles of which are discussed in [51]. This is relevant for the classification of data blocks that do not have clear markers. Many researchers use artificial intelligence methods for this purpose. Thus, classifiers based on support vector machines with handcrafted features have been used in previous studies [21 - 23, 28, 29, 31, 43, 44]. In these cases, the result of the identification of data blocks depended, among other things, on the correctness of the selection of classifier features. In more recent studies [25, 30, 32 - 37], support vector machines, k Nearest Neighbors and various types of neural networks with automatic feature extraction were applied. The above approach removed the human factor in selecting features and showed its suitability and high efficiency. Other works [26, 31, 33, 38] have made it possible to compare machine learning methods with each other. These studies show that using different types of neural networks to identify data blocks yields higher accuracy rates in most cases than other methods.

A comparison of the above methods showed that the achieved efficiency depends on the type of selected algorithm and the number of file types that were trained. In addition, the task is complicated by blocks of different data types in the compound files. As seen in [30, 32 - 35], using neural networks with automatic feature extraction is a perspective direction in data identification.

It should be noted that due to the wide variety of data types, some authors achieved prospective results in research on the identification of specific file types, such as AVI [24], audio [39], MPEG-4 and H.264 video formats [40], JPEG [45], PDF [46], SQLite [47], PNG [48], EVTX [59], and even compression algorithms [36]. These studies used advanced knowledge of the internal structure of these file types, which provided additional benefits in detecting and identifying such data.

After classifying fragments by data or file type, the next logical step is to perform clustering of these data blocks and file reconstruction. These tasks are closely intertwined and sometimes solved comprehensively. The case of bi-fragmented files is described in detail in [16]. The main problems appear with several file fragments and especially with the inconsistent placement of these data blocks.

In general, file reconstruction approaches are based on knowledge of the file's internal structure and/or content. For example, because of the complex structure of graphic files, various methods exist for recovering them. In [43, 44], we used the Parallel Unique Path algorithm (PUP), highlighted in [60]. On the other hand, to recover

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

graphic files, researchers have successfully proposed determining the similarity between pixel values [27, 55], comparing pixel values on the fragment boundaries [50], applying similarity metrics [45, 49], using PNG and JPEG internal structure features [48, 56], analyzing PRNU fingerprints of the cameras [54], and utilizing both internal structure and content of JPEG files [53]. In addition, the use of internal file structure for its recovery is possible with many types of compound files, such as video [40], SQLite databases [47], DOC [52], OOXML [57], BIN [58], and EVTX [59]. Instead, when recovering text documents, there is an additional option to use their content. Therefore, in these cases, it is possible to use dictionary-based techniques [4, 46, 50].

Noteworthy is the use of artificial intelligence techniques to restore audio [39] and graphic files [41] with damaged headers, as well as the use of a validator to reconstruct video files with lost areas containing video codec specifications [42]. In these papers, the authors proposed methods that provide access to the internal contents of damaged files. As seen from the above works [39, 41], artificial intelligence methods are a perspective direction in restoring media data content. In general, this can be seen as a way to replace computationally complex algorithms.

The analyzed works show that no universal tool can simultaneously solve all problems in the search, identification, and reconstruction of file fragments. As can be seen from Table 4, two tendencies are traced. In some cases (for instance, [23, 32, 33, 39]), researchers focus on creating new approaches or improving existing methods for specific stages of file carving. This mainly refers to the data identification phase. Because of the use of artificial intelligence at this stage, many approaches typically focus on identifying various file or data types, - up to 75 [30, 34, 35]. In other words, there is a certain universality in most cases.

Another tendency is to use the peculiarities of the internal structure of certain file types or their contents in file carving (for example, [4, 43, 46, 48]). The methods proposed in these papers are developed for identifying, clustering, reconstructing, or restoring only files of specific types. Almost each of these approaches (e.g., [47, 50, 57]) requires first studying the internal structure of a file type or gaining access to certain parts of its contents. Therefore, they are usually not appropriate for other file types.

Conclusions

This paper systematizes advanced file carving techniques and presents an ontological scheme of file carving. Although file carving techniques are generally known and understandable, they have several disadvantages when working with different types of fragmented files. As a result, many researchers have attempted to improve existing techniques and develop their own data recovery methods. The mentioned ontological scheme can be used as a roadmap for these purposes by digital forensics investigators.

At the beginning of the study, we identified three questions. The conclusions obtained from the analysis of the papers are summarized below.

Q1. What are the typical stages of file carving and what are the perspectives for improving each stage?

In general, in the case of data fragmentation, there is a tendency to divide the file carving process into stages to solve individual subtasks: 1) identification of data blocks without explicit markers and 2) classification and reconstruction of files or their contents.

The first of these stages, the identification of data blocks, is characterized by the widespread use of artificial intelligence techniques. Artificial intelligence models and methods have quite high efficiency. However, most researchers focus on identifying a limited range of data types. Therefore, a perspective direction is the development of models and methods that can identify a wide range of data block types and be self-learning. In addition, the analyzed techniques need to be improved to increase accuracy and prevent the loss of important data blocks in case of misclassification.

The main problems of the following phases are the difficulty clustering the detected data blocks, i.e., assigning a particular group of fragments to a specific file. Out-of-order fragmentation has additional issues with the correct assembly of the file. It can be concluded that there are no universal techniques at these stages, and all of them require a detailed analysis of the file types to be recovered.

Q2. Is it possible to carve fragmented files without priori information about their internal structure and contents?

The universal methods used to identify data blocks actually depend on the alphabet's power of the classification analysis models. At the same time, the reconstruction process of files depends on their internal structure and/or contents. Therefore, each described method is applied only to recover files of certain types. The only exception in some cases may be approaches for recovering bi-fragmented files.

Q3. What are the perspectives on using artificial intelligence methods in the field of file carving?

The role of artificial intelligence is not restricted to identifying data fragments. It is important to restore access to file contents in cases of overwriting or damaging some areas of files. Thus, artificial intelligence techniques are used to generate headers to restore the content of damaged media files. In general, artificial intelligence models and methods are a perspective approach to reduce complexity. Due to the universality of artificial

intelligence, it is possible to use artificial intelligence techniques to develop carving methods independent of the internal structure and content of files.

Limitations. This paper does not provide an overview of all available data recovery methods. Emphasis was placed on methods of recovering fragmented files with lost or damaged metadata. In addition, the goal was not to study methods of minimizing the cost of resources and time, such as building a map of unused data [61].

Future research should focus on increasing the accuracy and efficiency of the proposed methods and the resource and time economy. Improving artificial intelligence techniques for identifying blocks of data types will allow the detection of a more complete set of fragments of target file types and minimize erroneously omitted data. With regard to data reconstruction, due to the large variety of file types, the current issues are to improve existing methods and develop new approaches.

Contribution of authors: conceptualization of the problem, supervision and revision – **Viacheslav Moskalenko**; original draft preparation – **Maksym Boiko**; visualization, review, and editing – **Oksana Shovkoplias**.

All authors have read and agreed with the published version of this manuscript.

References

1. Carrier, B. *File System Forensic analysis*. Addison-Wesley Professional, 2005. 600 p.

2. Bonetti, G., Viglione, M., Frossi, A., Maggi, F., & Zanero, S. Black-box forensic and antiforensic characteristics of solid-state drives. *Journal of Computer Virology and Hacking Techniques*, 2014, vol. 10, no. 4, pp. 255–271. DOI: 10.1007/s11416-014-0221-z.

3. Ligh, M. H., Case, A., Levy, J., & Walters, A. *The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux, and Mac Memory 1st Edition.* John Wiley & Sons, 2014. 912 p.

4. Ali, N. U. A., Iqbal, W., & Afzal, H. Carving of the OOXML document from volatile memory using unsupervised learning techniques. *Journal of Information Security and Applications*, 2022, vol. 65, article no. 103096. DOI: 10.1016/j.jisa.2021.103096.

5. Darnowski, F., & Chojnaki, A. Selected methods of file carving and analysis of digital storage media in computer forensics. *Teleinformatics Review*, 2015, vol. 1-2, pp. 25–40. Available at: https://yadda.icm.edu.pl/ baztech/element/bwmeta1.element.baztech-10af3f4e-db53-4ae5-9b7f-b7e850dd08d0/c/Darnowski_F_Chojnacki_A.pdf (accessed 19.09.2023).

6. Pahade, R. K., Singh, B., & Singh, U. A Survey on Multimedia File Carving. *International Journal of Computer Science & Engineering Survey (IJCSES)*, 2015, vol. 6, no. 6, pp. 27–46. DOI: 10.5121/ijcses. 2015.6603.

7. Alrobieh, Z. S., & Raqpan, A. M. A. A. File Carving Survey on Techniques, Tools and Areas of Use. *Transactions on Networks and Communications*, 2020, vol. 8, no. 1, pp. 16–26. DOI: 10.14738/tnc.81.7636.

8. Al-Jawry, Rabei., & Mohamad, Kamaruddin., Jamel, Sapiee., & Ahmad Khalid, Shamsul Kamal. A review of digital forensics methods for JPEG file carving. *Journal of Theoretical and Applied Information Technology*, 2018, vol. 96, no. 17, pp. 5841-5856. Available at: http://www.jatit.org/volumes/Vol96No17/ 17Vol96No17.pdf (accessed 19.09.2023)

9. Rintu Aleyamma Thomas., & Mathai, M. A Survey on File Carving Process Using Foremost and Scalpel. *National Conference on Emerging Computer Applications (NCECA2021)*, Kerala, 2021, vol. 3, no. 1, pp. 70-72. DOI: 10.5281/ZENODO.5091663.

10. Ali, N. U. A., Iqbal, W., & Shafqat, N. Analysis of Windows OS's Fragmented File Carving Techniques: A Systematic Literature Review. *16th International Conference on Information Technology-New Generations (ITNG 2019).* Springer International Publishing, 2019, pp. 63–67. DOI: 10.1007/978-3-030-14070-0_10.

11. Sari, S. A., & Mohamad, K. M. A Review of Graph Theoretic and Weightage Techniques in File Carving. *Journal of Physics: Conference Series*. IOP Publishing, 2020, vol. 1529, no. 5. DOI: 10.1088/1742-6596/1529/5/052011.

12. Ramli, N. I. S., Hisham, S. I., & Razak, M. F. A. Survey of File Carving Techniques. *Innovative Systems for Intelligent Health Informatics (IRICT 2020)*. Lecture Notes on Data Engineering and Communications Technologies, Springer, 2021, vol 72, pp. 815–825. DOI: 10.1007/978-3-030-70713-2_74.

13. Alherbawi, N., Shukur, Z., & Sulaiman, R. A Survey on Data Carving in Digital Forensic. *Asian Journal of Information Technology*, 2016, vol. 15, no. 24, pp. 5137-5144. Available at: http://docsdrive.com/pdfs/ medwelljournals/ajit/2016/5137-5144.pdf (accessed 19.09.2023).

14. Kävrestad, J. Analyzing Data and Writing Reports. *Fundamentals of Digital Forensics*. Springer International Publishing, 2020, pp. 85–98. DOI: 10.1007/978-3-030-38954-3_10.

15. Lin, X. File Carving. *Introductory Computer Forensics*. Springer International Publishing, 2018, pp. 211–233. DOI: 10.1007/978-3-030-00581-8_9.

16. Garfinkel, S. L. Carving contiguous and fragmented files with fast object validation. *Digital Investigation*, 2007, vol. 4, pp. 2–12. DOI: 10.1016/j.diin.2007.06.017.

17. Dubettier, A., Gernot, T., Giguet, E., & Rosenberger, C. File type identification tools for digital investigations. *Forensic Science International: Digital Investigation*, 2023, vol. 46, article no. 301574. DOI: 10.1016/j.fsidi.2023.301574.

18. Alghafli, K., Jones, A., & Martin, T. Investigating and measuring capabilities of the forensics file carving techniques. *Future Information Technology*. Lecture Notes in Electrical Engineering, Springer, 2014,

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

vol 276, pp. 329–336. DOI:10.1007/978-3-642-40861-8_47.

19. Kloet, S. J. J. *Measuring and Improving the Quality of File Carving Methods*. MSc thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, The Netherlands, 2007. 111 p. Available at: https://research.tue.nl/files/46916835/635640 -1.pdf (accessed 25.06.2023)

20. Laurenson, T. Performance analysis of file carving tools. *IFIP Advances in Information and Communication Technology*. Security and Privacy Protection in Information Processing Systems, 2013, vol. 405, pp. 419–433. DOI: 10.1007/978-3-642-39218-4_31.

21. Zanero, S. File block classification by Support Vector Machines. *2011 Sixth International Conference on Availability, Reliability and Security*, Vienna, Austria, 2011, pp. 307-312. DOI: 10.1109/ARES.2011.52.

22. Fitzgerald, S., Mathews, G., Morris, C., & Zhulyn, O. Using NLP techniques for file fragment classification. *Digital Investigation*, 2012, vol. 9, pp.S44–S49. DOI: 10.1016/j.diin.2012.05.008.

23. Beebe, N. L., Maddox, L. A., Liu, L., & Sun, M. Sceadan: Using concatenated N-gram vectors for improved file and data type classification. *IEEE Transactions on Information Forensics and Security*, 2013, vol. 8, no. 9, pp. 1519-1530. DOI: 10.1109/TIFS.2013.2274728.

24. Pan, J., Liu, L., Sun, G., & Tang, Y. A method to identify the AVI-type blocks based on their four-character codes and C4.5 algorithm. *2014 International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC2014)*, Shanghai, China, 2014, pp. 1-7. DOI: 10.1109/BESC.2014.7059521.

25. Wang, F., Quach, T.-T., Wheeler, J., Aimone, J. B., & James, C. D. Sparse Coding for N-Gram Feature Extraction and Training for File Fragment Classification. *IEEE Transactions on Information Forensics and Security*, 2018, vol. 13, no. 10, pp. 2553-2562. DOI: 10.1109/TIFS.2018.2823697.

26. Karampidis, K., Kavallieratou, E., & Papadourakis, G. Comparison of Classification Algorithms for File Type Detection A Digital Forensics Perspective. *POLIBITS*, 2017, vol. 56, pp. 15-20. Available at: https://api.semanticscholar.org/CorpusID:51882719 (accessed 25.06.2023).

27. Al-Sadi, A., Yahya, M. B., & Almulhem, A. Identification of image fragments for file carving. *World Congress on Internet Security (WorldCIS-2013)*, London, UK, 2013, pp. 151-155. DOI: 10.1109/WorldCIS.2013.6751037.

28. Bhatt, M., Mishra, A., Kabir, M. W. U., Blake-Gatto, S. E., Rajendra, R., Hoque, M. T., & Ahmed, I. Hierarchy-Based File Fragment Classification. *Machine Learning and Knowledge Extraction*, 2020, vol. 2, no. 3, pp. 216-232. DOI: 10.3390/make2030012.

29. Sportiello, L., & Zanero, S. Context-based file block classification. *IFIP Advances in Information and Communication Technology*, 2012, vol 383, pp. 67-82. DOI: 10.1007/978-3-642-33962-2_5.

30. Mittal, G., Korus, P., & Memon, N. FiFTy: Large-Scale File Fragment Type Identification Using Convolutional Neural Networks. *IEEE Transactions on Information Forensics and Security*, 2021, vol. 16, pp. 28–41. DOI: 10.1109/TIFS.2020.3004266.

31. Sester, J., Hayes, D., Scanlon, M., & Le-Khac, N. A. A comparative study of support vector machine and neural networks for file type identification using n-gram analysis. *Forensic Science International: Digital Investigation*, 2021, vol. 36, article no. 301121. DOI: 10.1016/j.fsidi.2021.301121.

32. Chen, Q., Liao, Q., Jiang, Z. L., Fang, J., Yiu, S., Xi, G., Li, R., Yi, Z., Wang, X., Hui, L. C. K., Liu, D., & Zhang, E. File fragment classification using grayscale image conversion and deep learning in digital forensics. *2018 IEEE Security and Privacy Workshops (SPW)*, San Francisco, CA, USA, 2018, pp. 140-147. DOI: 10.1109/SPW.2018.00029.

33. Hiester, L. *File Fragment Classification Using Neural Networks with Lossless Representations*. Bachelor Thesis, East Tennessee State University. Undergraduate Honors Theses, 2018, Paper 454, 36 p. Available at: https://dc.etsu.edu/honors/454 (accessed 25.06.2023).

34. Ghaleb, M., Saaim, K., Felemban, M., Al-Saleh, S. M., & Al-Mulhem, A. File Fragment Classification using Light-Weight Convolutional Neural Networks. *arXiv (Cornell University)*, 2023. DOI: 10.48550/arxiv.2305.00656.

35. Liu, W., Wang, Y., Wu, K., Yap, K., & Chau, L. A Byte Sequence is Worth an Image: CNN for File Fragment Classification Using Bit Shift and n-Gram Embeddings. *arXiv (Cornell University)*, 2023. DOI: 10.48550/arxiv.2304.06983.

36. Bharadwaj, S. Using convolutional neural networks to detect compression algorithms. *arXiv (Cornell University)*, 2021. DOI: 10.48550/arxiv.2111.09034.

37. Haque, E., & Tozal, M. E. Byte embeddings for file fragment classification. *Future Generation Computer Systems*, 2022, vol. 127, pp. 448–461. DOI: 10.1016/j.future.2021.09.019.

38. Vulinovic, K., Ivkovic, L., Petrovic, J., Skracic, K., & Pale, P. Neural Networks for File Fragment Classification. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2019, pp. 1194-1198. DOI: 10.23919/mipro.2019.8756878.

39. Heo, H.-S., So, B.-M., Yang, I.-H., Yoon, S.-H., & Yu, H.-J. Automated recovery of damaged audio files using deep neural networks. *Digital Investigation*, 2019, vol. 30, pp. 117-126. DOI: 10.1016/j.diin.2019.07.007.

40. Na, G. H., Shim, K. S., Moon, K. W., Kong, S. G., Kim, E. S., & Lee, J. Frame-based recovery of corrupted video files using video codec specifications. *IEEE Transactions on Image Processing*, 2014, vol. 23, no. 2, pp. 517-526. DOI: 10.1109/TIP.2013.2285625.

41. Amrouche, S. C., & Salamani, D. Non-parametric adaptative JPEG fragments carving. *Tenth*

International Conference on Machine Vision, Vienna, Austria, 2017, article no. 106962D. DOI: 10.1117/12.2310079.

42. Alghafli, K., & Martin, T. Identification and recovery of video fragments for forensics file carving. *2016 11th International Conference for Internet Technology and Secured Transactions (ICITST)*, Barcelona, Spain, 2016, pp. 267-272. DOI: 10.1109/ICITST.2016.7856710.

43. Qiu, W., Zhu, R., Guo, J., Tang, X., Liu, B., & Huang, Z. A new approach to multimedia files carving. *2014 IEEE International Conference on Bioinformatics and Bioengineering*, Boca Raton, FL, USA, 2014, pp. 105-110. DOI: 10.1109/BIBE.2014.31.

44. Guo, J., He, J., & Huang, N. Research of Multiple-type Files Carving Method Based on Entropy. *Proceedings of the 2015 4th National Conference on Electrical, Electronics and Computer Engineering*, 2016, pp. 521-528. DOI: 10.2991/nceece-15.2016.98.

45. Ali, R. R., & Mohamad, K. M. RX_myKarve carving framework for reassembling complex fragmentations of JPEG images. *Journal of King Saud University - Computer and Information Sciences*, 2021, vol. 33, no. 1, pp. 21–32. DOI: 10.1016/J.JKSUCI. 2018.12.007.

46. Al-Sharif, Z. A., Al-Khalee, A. Y., Al-Saleh, M. I., & Al-Ayyoub, M. Carving and clustering files in RAM for memory forensics. *Far East Journal of Electronics and Communications*, 2018, vol. 18, no. 5, pp. 695 - 722. DOI: 10.17654/ec018050695.

47. Zhang, L., Hao, S., & Zhang, Q. Recovering SQLite data from fragmented flash pages. *Annals of Telecommunications*, 2019, vol. 74, no. 7–8, pp. 451–460. DOI: 10.1007/s12243-019-00707-9.

48. Hilgert, J. N., Lambertz, M., Rybalka, M., & Schell, R. Syntactical Carving of PNGs and Automated Generation of Reproducible Datasets. *Digital Investigation*, 2019, vol. 29, pp. S22-S30. DOI: 10.1016/j.diin.2019.04.014.

49. Tang, Y., Fang, J., Chow, K. P., Yiu, S. M., Xu, J., Feng, B., Li, Q., & Han, Q. Recovery of heavily fragmented JPEG files. *Digital Investigation*, 2016, vol. 18, pp. S108-S117. DOI: 10.1016/j.diin.2016.04.016.

50. Ravi, A., Kumar, T. R., & Mathew, A. R. A method for carving fragmented document and image files. *2016 International Conference on Advances in Human Machine Interaction (HMI)*, Kodigehalli, India, 2016, pp. 1-6. DOI: 10.1109/HMI.2016.7449170.

51. Roussev, V., & Garfinkel, S. L. File fragment classification - The case for specialized approaches. *2009 Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, Berkeley, CA, USA, 2009, pp. 3-14. DOI: 10.1109/SADFE. 2009.21.

52. Lin, W., & Xu, M. A Microsoft Word documents carving method base on interior virtual streams. *Advanced Materials Research*, 2012, vols. 433–440, pp. 3028-3032. DOI: 10.4028/www.scientific. net/AMR.433-440.3028.

53. Birmingham, B., Farrugia, R. A., & Vella, M. Using thumbnail affinity for fragmentation point detection of JPEG files. *IEEE EUROCON 2017 -17th International Conference on Smart Technologies*, Ohrid, Macedonia, 2017, pp. 3-8. DOI: 10.1109/EUROCON. 2017.8011068.

54. Durmus, E., Korus, P., & Memon, N. Every Shred Helps: Assembling Evidence from Orphaned JPEG Fragments. *IEEE Transactions on Information Forensics and Security*, 2019, vol. 14, no. 9, pp. 2372-2386. DOI: 10.1109/TIFS.2019.2897912.

55. Chang, X., Wu, J., & Hao, F. JPEG fragment carving based on pixel similarity of MED-ED. *2019 Chinese Control Conference (CCC)*, Guangzhou, China, 2019, pp. 8862-8866. DOI: 10.23919/ChiCC.2019. 8865161.

56. Uzun, E., & Sencar, H. T. JpgScraper : An Advanced Carver for JPEG Files. *IEEE Transactions on Information Forensics and Security*, 2020, vol. 15, pp. 1846-1857. DOI: 10.1109/TIFS.2019.2953382.

57. Boiko, M., & Moskalenko, V. Syntactical method for reconstructing highly fragmented OOXML files. *Radioelectronic and Computer Systems*, 2023, no. 1, pp. 166–182. DOI: 10.32620/reks.2023.1.14.

58. Hand, S., Lin, Z., Gu, G., & Thuraisingham, B. Bin-Carver: Automatic recovery of binary executable files. *Digital Investigation*, 2012, vol. 9, pp.S108–117. DOI: 10.1016/j.diin.2012.05.014.

59. Xu, M., Sun, J., Zheng, N., Qiao, T., Wu, Y., Shi, K., & Yang, T. A Novel File Carving Algorithm for EVTX Logs. *Digital Forensics and Cyber Crime. ICDF2C 2017*, Prague, Czech Republic, 2017, vol. 216, pp. 97–105. DOI: 10.1007/978-3-319-73697-6_7.

60. Memon, N., & Pal, A. Automated reassembly of file fragmented images using greedy algorithms. *IEEE Transactions on Image Processing*, 2006, vol. 15, no. 2, pp. 385-393. DOI: 10.1109/tip.2005.863054.

61. Karresand, M., Warnqvist, A., Lindahl, D., Axelsson, S., & Dyrkolbotn, G. O. Creating a Map of User Data in NTFS to Improve File Carving. *Advances in Digital Forensics XV. 15th IFIP WG 11.9 International Conference*, Orlando, FL, USA, 2019, pp. 133-158. DOI: 10.1007/978-3-030-28752-8_8.

Received 27.07.2023, Accepted 20.09.2023

УДОСКОНАЛЕНИЙ КАРВІНГ ФАЙЛІВ: ТАКСОНОМІЯ, МОДЕЛІ ТА МЕТОДИ

Максим Бойко, В'ячеслав Москаленко, Оксана Шовкопляс

Техніки карвінгу файлів мають важливе значення у сфері цифрової криміналітики. При цьому бурхливе зростання кількості і типів даних, обумовлює необхідність розвитку методів карвінгу файлів із точки зору

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

можливостей, точнісних характеристик та обчислювальної ефективності. Проте переважна більшість методів розробляється для вирішення конкретних вузьких задач і опирається на певний набір припущень і апріорних знань про файли, які потрібно відновити. Існує брак досліджень, що систематизують методи і структурують підходи задля виявлення прогалин і визначення перспективних напрямків розвитку з урахуванням останніх досягнень в галузі інформаційних технологій та штучного інтелекту. **Предметом** вивчення в статті є структура, фактори, критерії ефективності, методи та інструменти карвінгу файлів, а також поточний стан і тенденції розвитку методів карвінгу. **Метою** є систематизація знань про сучасні методи карвінгу файлів та виявлення перспективних напрямків розвитку. **Завдання**: виділити основні етапи карвінгу файлів і проаналізувати підходи до їх реалізації; побудувати онтологічну схему карвінгу файлів; визначити перспективні напрямки розвитку методів карвінгу файлів. Використовуваними **методами** є: літературний огляд, систематизація і узагальнення. Отримано такі **результати**. Побудовано онтологічну схему концепції карвінгу файлів. Схема включає в себе принципи, властивості, етапи, техніки, критерії оцінки, інструменти карвінгу файлів, а також фактори, що впливають на процес. Наведено особливості, обмеження та області застосування методів відновлення даних. Встановлено, що досі широкорозповсюдженим підходом до реконструкції файлів є ручне детальне вивчення внутрішньої структури файлів та/або їх вмісту, виявлення певних закономірностей, що дозволяють відтворити у правильному порядку послідовність фрагментів даних. При цьому переважна більшість методів не гарантує стовідсоткового результату. Проаналізовано поточний стан та перспективи використання методів штучного інтелекту в сфері комп'ютерно-технічної експертизи, зокрема для ідентифікації блоків даних, кластеризації та реконструкції файлів, а також відтворення вмісту медіафайлів з пошкодженими або втраченими заголовками. Визначено необхідність наявності апроіорної інформації про структуру або вміст файлів для успішності карвінгу фрагментованих даних. **Висновки**. Наукова новизна отриманих результатів полягає в наступному: вперше систематизовано і проаналізовано сучасні методи карвінгу файлів за напрямками розвитку і виявлено перспективність використання штучного інтелекту для ідентифікації блоків даних, кластеризації та відновлення вмісту файлів; вперше побудовано онтологічну схему карвінгу файлів, яка може бути використана як дорожня карта під час розроблення нових перспективних систем у сфері комп'ютерно-технічної експертизи.

Ключові слова: комп'ютерно-технічна експертиза; метадані; фрагментація; фрагментований файл; відновлення даних; карвінг файлів; ідентифікація фрагменту файлу; реконструкція файлу; відновлення файлу; штучний інтелект.

Бойко Максим Володимирович – асп. каф. комп'ютерних наук, Сумський державний університет, Суми, Україна; старший детектив, Управління аналітики та обробки інформації, Національне антикорупційне бюро України, Київ, Україна.

Москаленко В'ячеслав Васильович – канд. техн. наук, доц., доц. каф. комп'ютерних наук, Сумський державний університет, Суми, Україна; докторант каф. комп'ютерних систем, мереж та кібербезпеки, Національний аерокосмічний університет ім. М. Є. Жуковського "Харківський авіаційний інститут", Харків, Україна.

Шовкопляс Оксана Анатоліївна – канд. фіз.-мат. наук, старш. викл. каф. комп'ютерних наук, Сумський державний університет, Суми, Україна.

Maksym Boiko – PhD Student at Computer Sciences Department of Sumy State University, Sumy, Ukraine; Senior Detective, Information Processing and Analysis Department, the National Anti-Corruption Bureau of Ukraine, Kyiv, Ukraine,
e-mail: mboiko25@gmail.com, ORCID: 0000-0003-0950-8399, Scopus Author ID: 58199360000.

Viacheslav Moskalenko – PhD, Associate Professor at Computer Science Department of Sumy State University, Sumy, Ukraine; Doctoral Student at Department of Computer Systems, Networks and Cybersecurity, National Aerospace University "KhAI", Kharkiv, Ukraine,
e-mail: v.moskalenko@cs.sumdu.edu.ua, ORCID: 0000-0001-6275-9803, Scopus Author ID: 57189099775.

Oksana Shovkoplias – PhD, Senior Lecturer at Computer Science Department of Sumy State University, Sumy, Ukraine,
e-mail: o.shovkoplyas@mss.sumdu.edu.ua, ORCID: 0000-0002-4596-2524, Scopus Author ID: 55647364100.