Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 31, Num. 2 (2024) 110-119

RESEARCH ARTICLE

Analysis of Forensic Tools for Recovery of Formatted Data: a case study with Microsoft Word files

Análise de Softwares Forenses para Recuperação de Dados Formatados: um estudo de caso com arquivos do Microsoft Word

Rubens Karman Paula da Silva [1]*, Islan Amorim Bezerra [1], Sidney Marlon Lopes de Lima [2], Sérgio Murilo Maciel Fernandes [1]

Abstract: Deleting or formatting files in order to conceal a crime can be frustrated because of how easy it is to use forensic software that use data carving techniques. This research wants to evaluate the accuracy of forensic data carving software when tested to recover formatted Microsoft Word files. The chosen software is widely used in the field and has been featured in scientific papers such as Foremost, Scalpel, Recurva, PhotoRec, Autopsy and Magic Rescue. The metrics taken into consideration were: software execution time, number and size of files recovered, number of false and true positives generated in three test scenarios. The study was validated by comparing the resulting files with the originals using a hash algorithm. To create the test scenarios, a dataset was built with 16,000 copies of files of various lengths. In each scenario, there was a difference between the number of files and the requirements that the software was tested, with only doc or docx files being supposed to be recovered. Among the software analyzed, Recuva, Autopsy and PhotoRec had the highest percentage of true positives (90%) in all the scenarios tested. When it comes to false positives, Recuva performed better than the others, with approximately 1%.
Keywords: Digital forensics — Data Carving — File Carving — Performance analysis — Word documents

Resumo: Excluir ou formatar arquivos para esconder ações ilícitas pode ser considerado uma ação frustrada, considerando a facilidade de utilizar softwares forense que implementam técnicas de data carving. A presente pesquisa propõe avaliar a acurácia de softwares forense de data carving quando sujeitados a recuperar arquivos do Microsoft Word formatados. Os softwares escolhidos são aplicações amplamente difundidas na área e possuem destaque em trabalhos científicos, sendo eles: Foremost, Scalpel, Recurva, PhotoRec, Autopsy e Magic Rescue. As métricas analisadas foram: tempo de execução do software, quantidade e tamanho dos arquivos recuperados, quantidade de falsos positivos e verdadeiros positivos gerados em três cenários de testes. A validação ocorreu comparando os arquivos resultantes com os originais através de algoritmo hash. Para estruturar os cenários de testes, foi construído um dataset com 16.000 exemplares de arquivos de diversas extensões. Em cada cenário, variou-se a quantidade de arquivos e exigências que os softwares foram submetidos, sendo compelidos a recuperar apenas arquivos doc ou docx. Dos softwares analisados o Recuva, Autopsy e PhotoRec apresentaram maiores percentuais de verdadeiros positivos (>90%) em todos os cenários avaliados. Quanto a falsos positivos, o Recuva apresentou melhor resultado que os demais, com aproximadamente 1%.
Palavras-Chave: Forense digital — Data Carving — File Carving — Análise de desempenho — Documentos do Word

[1] *Departamento de Engenharia da Computação, Universidade de Pernambuco, Recife, Pernambuco, Brasil*
[2] *Departamento de Eletrônica e Sistemas, Universidade Federal de Pernambuco, Recife, Pernambuco, Brasil*
***Corresponding author**: rkps@ecomp.poli.br
DOI: https://doi.org/10.22456/2175-2745.140149 • **Received:** 12/05/2024 • **Accepted:** 10/07/2024

1. Introdução

A Forense Digital oferece um arcabouço metodológico e ferramental para elucidar casos de dano material, físico ou à dignidade de terceiros através de dispositivos digitais. Ações deixam vestígios, independentemente da intenção das pessoas. Sejam elas bem ou mal intencionadas, esses vestígios podem

ser rastreados e trazidos à luz do esclarecimento através dessa ferramenta.

Restaurar arquivos perdidos é uma técnica de fundamental relevância para a perícia forense digital, auxiliando peritos, investigadores e profissionais ao fornecer novos pontos de vista e *insights* quando aplicada à recuperação de dados formatados. É definindo como o processo pelo qual dados excluídos ou inacessíveis armazenados em mídias computacionais são recuperados [1, 2]. A técnica forense para alcançar esses registros na mídia é conhecida como *data carving*.

A abordagem do *data carving* é um importante tópico existente na área da Forense Digital. Sendo considerada um processo de aquisição, autenticação, análise e documentação de evidências contidas em um sistema digital [3].

O presente trabalho tem interesse particular em arquivos com extensão .doc e .docx do processador de texto Microsoft Word, uma vez que são amplamente utilizados em ambientes corporativos e residenciais. Esses arquivos são frequentemente usados para redigir e organizar informações, incluindo relatórios, contratos, comunicações internas, e outros documentos importantes [4]. Seu conteúdo pode ser de grande valor em processos judiciais, pois muitas vezes contém informações e evidências relevantes para a resolução de litígios. A integridade e autenticidade desses arquivos são, portanto, essenciais para garantir a justiça e a qualidade na análise forense digital. Conforme apontado por Hanis et al.[4], 2021, a análise meticulosa desses documentos pode revelar detalhes importantes sobre a autoria, possíveis tentativas de adulteração ou falsificação, fundamentais para a investigação e para a tomada de decisões judiciais.

Portanto, para responder à pergunta de pesquisa: entre os *softwares* de recuperação de arquivos formatados, comerciais e não comerciais, quais apresentam melhor acurácia, ou seja, menor percentual de falsos positivos, quando especificamente orientados a recuperar arquivos doc/docx? Foram realizados diversos experimentos em três cenários de testes distintos, fazendo uso de um *dataset* construído pelos autores. Ao final, os *softwares* de recuperação *Recuva, Autopsy* e *PhotoRec* apresentaram maiores percentuais de verdadeiros positivos ($>$90%) em todos os cenários avaliados.

1.1 Objetivo e contribuições do trabalho

Uma vez que as mídias de armazenamento estão cada vez maiores e ubíquas, uma análise pericial de dados formatados pode gerar uma enorme quantidade de dados que o perito deve analisar. Sabendo quais *softwares* pode gerar mais dados corrompidos (falsos positivos), o perito terá consciência de qual escolher mediante sua necessidade. Isso oportuniza uma otimização do seu trabalho, permitindo que ele se dedique ao que de realmente importa: a busca por ilicitudes no objeto periciado.

A presente pesquisa tem como principal objetivo a avaliação de desempenho de *softwares* de recuperação de dados formatados. A intenção é verificar o quanto cada *software* gera de: (*i*) falsos positivos, (*ii*) verdadeiros positivos e (*iii*) verdadeiros positivos duplicados de arquivos do tipo documentos em uma perícia forense.

Descrevemos os fundamentos teóricos básicos relacionados à abordagem de *data carving*, bem como suas limitações e trabalhos relacionados na seção 2. Na seção 3, é apresentado o processo metodológico, incluindo o *dataset*, métricas e cenários avaliados. A seção 4 mostra os resultados de cada estudo de caso. Por fim, a seção 5 com conclusões e trabalhos futuros.

2. Fundamentação Teórica

Na abordagem tradicional de aplicações *data carving*, que consiste em fazer uso do método de extração de arquivos baseados no padrão de assinatura de início e fim de um arquivo. Durante a extração, ao encontrar um cabeçalho conhecido, varre-se o sistema de arquivo continuamente até encontrar o fim do arquivo (rodapé). Os dados situados entre esses dois pontos (cabeçalho e rodapé) são extraídos e classificados como arquivo recuperado [3]. Para esse tipo de abordagem, os *softwares* de *data carving*, comumente conhecidos na literatura, como *Foremost*, *Photorec* e outros, utilizados nesta pesquisa, são empregadas a referida abordagem. Esses podem ser descritos como *carving* sequenciais, visto que seu algoritmo segue as diretivas descritas anteriormente [5].

Para cada extensão de arquivo existente, há uma sequência de caracteres em hexadecimal que definem seu cabeçalho e rodapé. Os arquivos de extensão PDF, por exemplo, têm a assinatura inicial, o cabeçalho, começando da mesma forma. A Tabela 1, adaptada de Povar e Bhadran[3], 2011, exemplifica uma sequência de assinaturas de cabeçalhos e rodapés suportadas pelo software do autor. Importante perceber que, no trabalho do autor, alguns cabeçalhos e rodapés não possuem sequência hexadecimal conhecida.

Os arquivos JPEG, por sua vez, têm em sua estrutura o cabeçalho hexadecimal ”0xFFD8” e o rodapé ”0xFFD9”, como mostra a Figura 1 construída pelos autores. Isso permite diferenciá-los de outros modelos de arquivos, com base em um exame do conteúdo.

A pesquisa de sequência de cabeçalho e rodapé também é conhecida como “números mágicos”. É importante atentar que, se a sequência de rodapé não for identificada na partição periciada, o arquivo não poderá ser recuperado. o que podendo ser uma limitação. Não finalizar o arquivo no momento certo pode comprometer a integridade do mesmo.

A abordagem *data carving* apresenta algumas limitações que restringem seu uso em determinadas situações. A primeira limitação diz respeito aos arquivos fragmentados. Esse tipo de programa não pode recuperar arquivos que tenham sido fragmentados em diferentes *clusters*. Ainda no tocante às limitações, esses *softwares* não validam muito bem o que é resultante da extração. Em decorrência disto, esse tipo de método apresenta uma quantidade muito grande de falsos positivos, ou seja, arquivos que efetivamente não existiam na mídia digital periciada original [6].

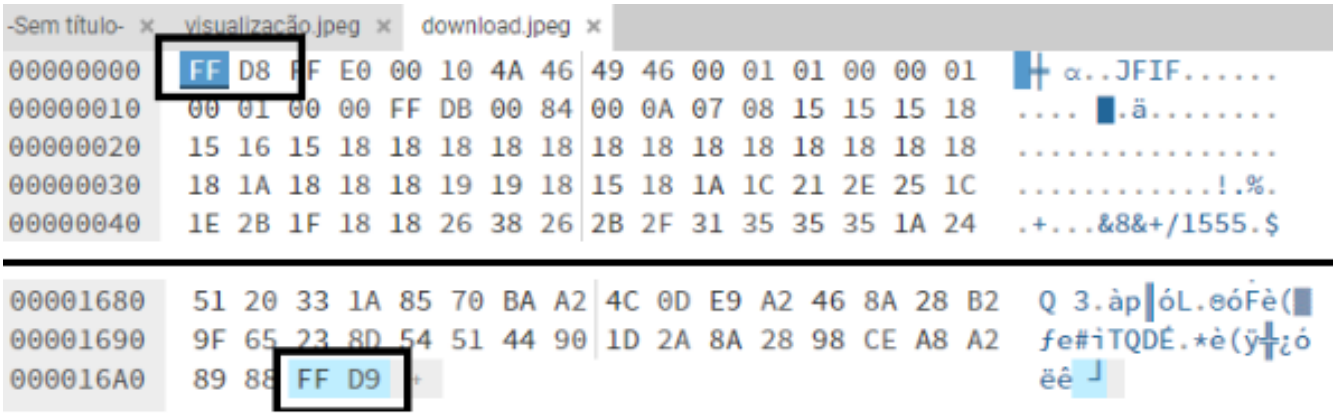

Figure 1. Cabeçalho e rodapé em hexadecimal de um arquivo JPEG. Fonte: Autoria própria

Table 1. Assinaturas de cabeçalhos e rodapés. Adaptado de Povar e Bhadran[3], 2011.

File	Header signature	Footer signature / Method of carving
JPEG	FFD8	FFD9
GIF	47494638	003B
PNG	89504E470D0A1A0A	49454E44
HTML	3C48544D4C3E	3C2F68746D6C3E
PDF	25504446	2525454F46
DOC	D0CF11E0A1B11AE1	File structure based carving
ZIP	504B0304	-
BMP	424D	File size is embedded in the header
AVI	52494646	-
MP4	66747970	File structure based carving
WMV	3026B2758E66CF11	-

Outras abordagens de implementação da técnica de *data carving* disponíveis na literatura são denominadas baseadas no conteúdo e baseadas na estrutura. A técnica baseada em conteúdo consiste em examinar cada cluster individualmente no conjunto de dados e analisar seu conteúdo para identificar relações entre os clusters que pertencem a um determinado arquivo já conhecido. Segundo Laurenson[7], 2013, isso envolve a obtenção de informações de metadados, como contagens de caracteres ou estatísticas sobre os bytes dos clusters. Por sua vez, na abordagem baseada na estrutura, este método inicia determinando o nível específico de informações sobre o formato do arquivo de interesse. Em seguida, essas informações são comparadas com o conjunto de dados brutos para reconhecer um arquivo [8]. Ambas as técnicas são encontradas na literatura em trabalhos que aplicam sobre dados fragmentados.

2.1 Limitações do *Data Carving*: *cluster* não contínuos

Um *cluster*, ou unidade de alocação, consiste em uma unidade básica de alocação de armazenamento em disco. De acordo com Silberschatz, Galvin e Gagne[9], 2018, um *cluster* corresponde a um grupo de setores de disco de potência 2. Nos sistemas de arquivos modernos, seu tamanho pode variar conforme configurado pelo usuário, mas geralmente são definidos entre 512 *bytes* e 64 KB.

Durante o processo de leitura e escrita em um sistema de armazenamento, quando o sistema de arquivos aloca os dados de forma contígua, os *clusters* são preenchidos sequencialmente conforme o momento são gravados no disco. À medida que os arquivos são persistidos e excluídos no disco durante o uso do SO, ocorrem casos considerados como fragmentação interna. A fragmentação é caracterizada quando os arquivos não são armazenados de forma contínua no *cluster*. Entre um arquivo e outro, e até mesmo dentro de um arquivo, podem existir *clusters* desconexos [10]. Assim, à medida que os dados são inseridos, modificados e deletados, formam-se *clusters* "vazios" que se tornam fragmentados.

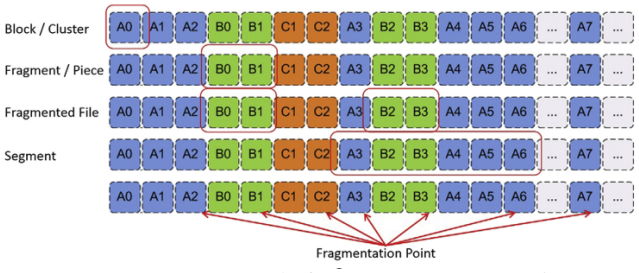

Figure 2. Demonstração de fragmentação em arquivos. [11]

A Figura 2, descrita por Tang et al.[11], 2016, ilustra claramente a fragmentação de arquivos que estão armazenados em *clusters*. Cada arquivo distinto é representado por uma cor e uma sequência de letras e números. Por exemplo, o arquivo colorido em azul possui uma sequência de 3 *clusters* (A0..A2). Posteriormente, ele é fragmentado, 4 *clusters* após é continuado por mais 1 *cluster* (A3), ocorrendo outra fragmentação e,

2 *clusters* após, o arquivo é continuado (A4...).

2.2 Limitações do *Data Carving*: Falsos Positivos

Outro ponto a ser considerado é que a abordagem de *data carving* não realiza rotinas de validação sobre o que está sendo recuperado de sua base. Em decorrência dessa limitação, a técnica resultar em muitos falsos positivos, ou seja, uma quantidade expressiva de arquivos recuperados, mas contém informações inválidas, faltantes ou que não existiam [6].

Segundo Laurenson.[12], 2013, os arquivos recuperados classificados como verdadeiros positivos são arquivos corretos recuperados do dispositivo digital periciado. Já os falsos positivos são arquivos classificados como incorretos ou corrompidos, ou seja, não são verdadeiros positivos. Por fim, arquivos falsos negativos são partes de um arquivo que não foram recuperados adequadamente. Quanto maior a quantidade de falsos positivos retornado de uma fonte periciada, maior o esforço demandado pelo perito e, consequentemente, mais tempo será preciso para concluir a investigação. Em determinadas situações, a análise forense pode ficar comprometida ou inviável.

Softwares considerados estados da arte, amplamente utilizados em ambientes de produção ou como comparativos em pesquisas científicas, naturalmente geram uma quantidade acentuada de falsos positivos. O *Scalpel*, por exemplo, em seu arquivo de configuração, informa previamente ao usuário sobre diretivas que podem amplia a quantidade de falsos positivos gerados. A importância de *softwares* que atuem na mitigação desses problemas, como proposto por Glaydyshev e James[5], 2017, que propõem uma abordagem de teoria da decisão para alcançar resultados com menor esforço sob determinadas circunstâncias, torna-se imprescindível.

2.3 Trabalhos relacionados

Em seu trabalho, Nurhayati e Fikri[13], 2017 realiza uma análise comparativa entre dois *softwares* de *data carving*: *PhotoRec* e *Foremost*. Em sua proposta de método experimental, o autor propõe oito cenários de testes que consideram cinco etapas importantes: preparação do dispositivo de armazenamento; criação da imagem do dispositivo (.dd ou .raw); *softwares* de recuperação de arquivos, (*PhotoRec e Foremost*); os arquivos de saída dos *softwares* de *carving*, ou seja, os arquivos previamente existentes anteriormente no dispositivo de armazenamento; e a validação dos arquivos recuperados. O estudo conclui que o *PhotoRec* apresenta melhor desempenho em termos de velocidade de processamento. Apesar de recuperar menos arquivos que seu concorrente, o *PhotoRec* retorna mais verdadeiros positivos. O referido trabalho serviu como norteador metodológico e experimental para a pesquisa atual.

Considerando o sistema de arquivo NTFS, Karresand, Dyrkolbotn e Axelsson.[14], 2020 tem o objetivo de aprofundar o conhecimento no referido sistema de arquivo em sistemas operacionais Windows, bem como o comportamento de sua integridade ao longo do tempo. Problemática relevante para a forense digital, uma vez que o Windows tem o maior *market share* do mercado de sistemas operacionais, sendo largamente utilizado por usuário final. O autor conclui que o comportamento de alocação de arquivos difere entre versões do *Windows* usando NTFS e que muda com o tempo, à medida que o sistema de arquivos cresce. Portanto, é importante conhecer os aspectos técnicos inerentes ao sistema de arquivo periciado.

Em sua pesquisa, Wu e Breitinger[15], 2020 realiza uma revisão da literatura com 799 trabalhos publicados entre 2014 e 2019. As perguntas da pesquisa estão relacionadas aos *softwares* forense em diferentes contextos de aplicação disponíveis de terceiros. Este estudo é relevante para nossa pesquisa, uma vez que ela expande as subáreas da forense digital, situando as técnicas de data carving dentro do conjunto denominado *multimedia forensics*. O único *software* carving apontado pelo estudo é o *"DECA: a decision-theoretic carving program"* de autoria de Glaydyshev e James[5], 2017.

Com maior frequência, na literatura são encontrados *softwares carving* voltados para arquivos de imagem. Por exemplo, Uzun e Sencar[16], 2020 propõem o JpgScraper. Segundo os autores, o *software* é uma solução inovadora para detectar dados JPEG, considerando *bitstream*, quantização, largura da imagem e outras características, a fim de recuperar arquivos órfãos fragmentados. No trabalho de Hilgert, Lambertz e Rybalka[17], 2019, é proposto um método de *data carving* que explora ao máximo a sintaxe de formato dos arquivos PNG. Essa abordagem alcança resultados de recuperação de 98% de arquivos PNG considerando, inclusive, cenários complexos de arquivos fragmentados.

Em contraste com a abundância de trabalhos científicos que lidam com imagens, softwares aplicados à recuperação de arquivos do tipo documentos são menos comuns. Tratado nos trabalhos de Hanis et al.[4], 2021 e Ravi, Kumar e Mathew[18], 2016, os autores utilizam abordagens mais avançadas do que os *magic numbers* para a recuperação de dados. São empregados algoritmos de árvore de decisão e métodos estatísticos, respectivamente. Hanis et al.[4], 2021 propõem recuperar documentos nos formatos PDF, DOC, DOCX, RTF e TXT, destacando a dificuldade natural dessas extensões em serem recuperados, devido à complexidade de encontrar o rodapé e, consequentemente, a finalização do arquivo. A obra propõe análise de idiomas (inglês, chinês e persa) no contexto de identificação do tipo de documento fragmentado. Por sua vez, Ravi, Kumar e Mathew[18], 2016 empregam dicionários para detectar fragmentos de arquivos .txt que estejam adjacentes ao *cluster* buscado, sendo necessário utilizar dicionários específicos para cada idioma.

Pereira et al.[19], 2019, em revisão da literatura, apontaram um panorama muito particular da ausência de trabalhos que abordem os falsos positivos em *softwares* de *Data Carving*. Eles evidenciaram que, entre os 107 trabalhos analisados, apenas um único trabalho abordou esse aspecto. Revelando uma carência de cuidados experimentais na concepção e implementação dos *softwares*, que não têm como meta o tratamento deste tipo de problema. Portanto, conscientes do

gap identificado na literatura apontado pelos autores e suas propostas de trabalhos futuros, os quais sugerem uma melhor compreensão das ferramentas mais utilizadas no processo de recuperação de dados, foi proposta a presente pesquisa.

3. Procedimento Metodológico

Neste capítulo serão descritos os percursos metodológicos realizados na pesquisa, abrangendo desde o processo de limpeza da mídia até a validação dos arquivos recuperados. A pesquisa é de caráter experimental, definida por Prodanov e Freitas[20], 2013. Trata-se de investigações cujos cujo objetivo principal é testar hipóteses que envolvem relações de causa e efeito. Ela emprega técnicas estatísticas na manipulação controlada das variáveis independentes, permitindo a observação de tais relações.

A Figura 3 ilustra as etapas metodológicas realizadas pelos autores para condução da pesquisa. Iniciando com a limpeza da mídia (*pendrive*) utilizada até a análise dos arquivos recuperados pelas aplicações *benchmark*. Nas subseções a seguir, serão descritos com maiores detalhes os procedimentos realizados em cada estágio.

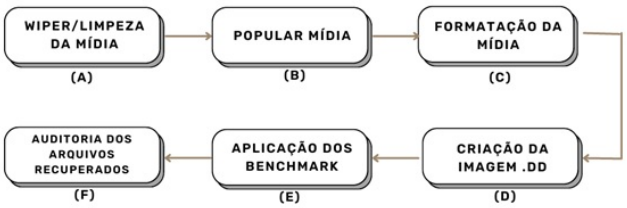

Figure 3. Etapas experimentais da pesquisa. Fonte: Autoria própria

3.1 Limpeza, anonimização e preparação de mídia

Como um dos preceitos de melhores práticas, elencados por Ali[21], 2012, o papel do perito é obter provas relacionadas ao caso e apresentá-las de forma admissível em um tribunal de justiça. Cabe ao perito realizar cópias exatas (*bit-a-bit*) das evidências originais.

O trabalho de análise deve ser realizado nas cópias, enquanto os originais permanecem íntegros, preservados e livres de qualquer alteração ou destruição. Para simular esse procedimento em uma mídia (*pendrive* de 8GB), foi realizado previamente o processo de *wiper* (limpeza) com o *software* de antiforense Shred. Este procedimento consiste em sobrescrever as trilhas e setores do sistema de arquivo da mídia desejada, apagando permanentemente qualquer dado que estiver armazenado.

3.2 Base de dados utilizada

Para atingir o objetivo da pesquisa, foi necessário a construção de um *dataset* que serviu como base de dados para a realização dos experimentos. Este *dataset* consiste em um diretório contendo 16.000 arquivos de diferentes extensões. As extensões coletadas são variadas e contemplam textos, planilhas, imagens, e-mails e banco de dados, além de representarem softwares proprietários e *open source*, como é possível verificar na Tabela 2.

Todos os arquivos existentes no *dataset* foram coletados da internet ou criados pelos próprios autores que, totalizando 5,1 GB. A base de arquivos, a cópia forense das partições utilizadas e os resultados obtidos estão disponíveis para consulta no presente link.

Para validar a base de dados, todos os arquivos foram submetidos a uma análise de semelhança para verificar se todos eram únicos. Essa validação consistiu na verificação de autenticidade através do número *hash* de cada arquivo, além de uma busca detalhada por duplicatas. Caso identificada uma duplicata, o arquivo repetido seria excluído, garantindo assim a ausência de arquivos idênticos. Posteriormente, cada arquivo na base foi renomeado com seu respectivo número *hash*. O objetivo é evitar duplicidade de nomes e ocultar qualquer informação que possibilite identificação de autoria do arquivo.

A etapa de Popular Mídia, Figura 3 (b), consiste em transferir do *dataset* para mídia, simulando um objeto suspeito que será periciado. Após o preenchimento, a mídia é apagada por meio do software de formatação padrão do Windows, Figura 3 (c). Consiste na tentativa de replicar o comportamento do suspeito, intencional ou não, de deletar informações comprometedoras.

A próxima etapa, Figura 3 (d), iniciou-se com a aplicação das boas práticas da perícia forense computacional. Nenhuma análise foi realizada diretamente na mídia periciada; em vez disso, foi feita uma cópia bit-a-bit e toda a análise foi conduzida nessa cópia. A presente pesquisa segue a recomendação do Simpósio Brasileiro de Segurança da Informação e emprega o software dd (*disk dump*), nativa em distribuições Linux, para a cópia forense.

3.3 *Benchmark*

Os softwares utilizados como *benchmark* são *softwares* comerciais tradicionalmente utilizados em forense digital para recuperar dados formatados. Os *softwares* utilizados na pesquisa são: (*i*) *Recurva*, (*ii*) *Foremost*, (*iii*) *Scalpel*, (*iv*) *MagicRescue*, (*v*) *PhotoRec* e (*vi*) *Autopsy*.

Eles são citados em diversos trabalhos com a finalidade de servir como parâmetro comparativo para novas implementações. Fazem uso as seguintes obras: Stanković e Khan[1], 2022; Glaydyshev e James[5], 2017; Wei, Zhen e Xu.[6], 2010; Laurenson.[12], 2013; Hilgert, Lambertz e Rybalka[17], 2019; Schneider, Eichhorn e Freiling[22], 2022; Birmingham, Farrugia e Vella[23], 2017; Palmieri e Zargari.[24], 2017 e Nurhayati e Fikri[13], 2017.

3.4 Auditoria e Métricas

Na etapa de "Auditoria dos arquivos recuperados", Figura 3 (e), as métricas de avaliação de desempenho dos *softwares* forense consideradas: a quantidade de falsos positivos gerados, o quantitativo de dados íntegros (verdadeiros positivos) e o

Table 2. Extensões de arquivos do dataset

Classe	Contra-classe	Descrição
JPEG	PNG	Correspondem a formatos de imagens.
Word (doc/docx)	Write – LibreOffice (odt)	Editores de texto utilizados para criar artigos, documentos, entre outros.
Excel (xls/xlsx)	Calc – LibreOffice (ods)	Editores de planilhas eletrônicas, utilizadas para gerenciar dados e cálculos matemáticos.
Power Point (ppt/pptx)	Impress – LibreOffice (odp)	Utilizado para criação gráfica e apresentações de slides.
Access (accdb)	SQL	Utilizados para criação e gerenciamento de banco de dados.
Outlook (msg)	Gmail (eml)	Extensões de emails.
OneNote (one)	Joplin (jex)	Utilizados como caderno virtual.
Publisher (pub)	TIFF	Utilizado para criação de cartões, *fly*, entre outros arquivos visuais.

tempo de execução dos *softwares* durante o processo. Com o intuito de validar os resultados recuperados de cada software, foi necessário a construção de um *script* para essa finalidade.

O *script* funciona da seguinte forma: obtém-se o código *hash* dos arquivos recuperados dos softwares; em seguida, é comparado com o código *hash* com o código de cada arquivo na base de dados; se houver correspondência, é considerado que o software conseguiu recuperar um arquivo íntegro (verdadeiro positivo), então o *script* registra como semelhante; os arquivos que passam pela verificação e não têm correspondência na base de dados são considerados falsos positivos. Também foi analisada a quantidade de verdadeiros positivos repetidos, visto que os *softwares* podem recuperar, de forma mais desnecessária, um mesmo arquivo incontáveis vezes.

3.5 Recursos computacionais utilizados

Para a execução dos experimentos propostos nessa pesquisa foi utilizado um computador com a seguinte configuração: CPU Intel i5-7200U 2.50GHz, 8GB de Memória RAM DDR4 1066 MHz, SSD 1GB e Nvidia GeForce 920MX DD3 2GB.

4. ESTUDO DE CASO E RESULTADOS

Os testes de validação foram separados em três cenários distintos:

4.1 Estudo de Caso: Cenário 1

No primeiro caso, o qual chamaremos de Cenário 1, considerado mais elementar devido à baixa quantidade de arquivos que os *softwares* de recuperação foram submetidos. Nesse primeiro momento, a atenção foi direcionada a apenas arquivos do tipo .doc e .docx por sua ampla utilização em ambientes corporativos e residenciais. Foram retirados 1.000 arquivos com essa extensão, os quais totalizaram tamanho de 66 MBs, e, já submetidos ao procedimento da seção 3.2, carregados na mídia. Em seguida foi realizada a formatação da mídia e criação da imagem bit-a-bit. A imagem resultante foi submetida a cada um dos *softwares* de recuperação de arquivos. A Tabela 3 contém o resultado de cada um dos *softwares*, bem como as métricas propostas para avaliação dos mesmos.

No critério de arquivos recuperados, é digno de destaque a precisão dos resultados auferidos com o Recurva. Foram recuperados exatamente a mesma quantidade de arquivos, com tamanho idêntico aos arquivos da base de dados, alcançando uma acurácia de 99% na verificação de que esses arquivos são realmente os originais. Outros *softwares* que merecem destaque devido à baixa taxa de falsos positivos são o *PhotoRec* e *Autopsy*. Ambos apresentaram 99% de verdadeiros positivos. Porém, é importante notar que o *PhotoRec* não conseguiu recuperar 229 arquivos existentes na mídia original, recuperando proporcionalmente 47,7 MB dos arquivos. Em contrapartida, o *Autopsy* recuperou 743 arquivos além do esperado na imagem *bit a bit*, mesmo estes sendo arquivos verdadeiros. Do resultado de verdadeiros positivos, 42,4% são arquivos duplicados, o que indica que foram geradas cópias de arquivos presentes na base original, acarretando maior quantidade de recuperados e maior ocupação de espaço em disco.

O *Scalpel*, *Foremost* e *Magic Rescue* foram os que menos atenderam às expectativas em seus resultados. Esses *softwares* recuperaram tanto arquivos a mais quanto a menos, com percentual de falsos positivos em demasia. No pior caso entre todos os *softwares*, podemos apontar o *Scalpel*, além de ter recuperado 14,5 GB de arquivos, gerou mais de 25 vezes o número de arquivos existentes na base de dados. Em resumo, todos os arquivos resultantes foram incorretos, visto que o percentual de falsos positivos foi de 100

No quesito tempo, o *softwares* mais ágil para entregar os arquivos formatados foram Foremost, Magic Rescue e Recurva, respectivamente. Porém, o Recurva apresentou melhor acurácia em comparação aos outros. É valido ressaltar que o *Autopsy*, durante o processo de recuperação, exigiu mais tempo para recuperação de metadados dos arquivos formatados. Caso não houvesse esse procedimento, o tempo de recuperação seria de aproximadamente 13m e 45s.

4.2 Estudo de Caso: Cenário 2

O presente estudo de caso, denominado Cenário 2, consiste em um experimento mais robusto, com quantidade e diversidade maior de arquivos, de forma proposital a analisar os *softwares* submetidos a uma carga elevada de dados. Foi utilizado o *dataset* descrito na seção 3.2 que possui 16.000 arquivos distribuídos em 16 extensões distintas (1.000 exemplares de

Table 3. Cenário 1: Resultados dos softwares periciados

Softwares	Qtd. de arquivos gerados	Tempo de Execução	Tamanho do diretório recuperado	Falsos positivos (%)	Verdadeiros positivos (%)	Verdadeiros positivos repetidos (%)
Recuva	1000	8:22	64 MB	<1%	>99%	0%
Foremost	1563	4:16	191,2 MB	38,7%	61,3%	0%
Scalpel	25077	1:37:24	14,5 GB	100%	0%	0%
Magic Rescue	162	8:10	56,8 MB	94%	6%	0%
Photorec	768	20:53	47,7 MB	<1%	>99%	0%
Autopsy	1743	40:51	112 MB	<1%	>99%	42,4%

cada extensão), totalizando um tamanho de armazenamento de 5,1 GBs.

Após os procedimentos metodológicos descritos na seção 3, a imagem *bit-a-bit* foi submetida aos *softwares* de recuperação de dados formatados. É apresentada na Tabela 4 os resultados dos *softwares* submetidos à avaliação neste Cenário 2. Neste experimento, ao analisar os resultados da métrica de verdadeiros positivos, os *softwares* que obtiveram melhores acurácias foram *Recuva*, *PhotoRec* e *Autopsy*, respectivamente.

Um destaque significativo vai para o *PhotoRec*, visto que o *software* apresentou melhores resultados em todas as métricas analisadas: tempo de execução, quantidade de arquivos recuperados mais próxima do total original (16.000) e tamanho total dos arquivos recuperados. Apesar de não ter conseguiu recuperar aproximadamente 1.170 originais, teve desempenho melhor do que o *Recuva* que, por sua vez, deixou de recuperar 2.048 originais.

Os resultados do *Autopsy* apresentam uma característica que exige reflexão. O *software* recuperou mais de 32 mil arquivos, dos quais quase 29 mil oriundos dos arquivos originais, ou seja, verdadeiros positivos, enquanto na base de dados original existiam apenas 16 mil arquivos. Supõe-se que na etapa de recuperação, o *software* recuperou cópias repetidas de arquivos originais.

Caso semelhante ao ocorrido no Cenário 1, mesmo tento um alto percentual de verdadeiros positivos, foi constatado um percentual de 44,5% de verdadeiros positivos repetidos. O tamanho excepcional desses arquivos é causado devido aos falsos positivos, uma vez os verdadeiros positivos são semelhantes aos arquivos originais e esses, por sua vez, possuíam aproximadamente 5,2GB.

Dos *softwares* apresentados, os três que obtiveram resultados insuficientes foram *Foremost*, *MagicRescue* e *Scalpel*, devido à sua alta taxa de falsos positivos. Mesmo gerando uma quantidade de arquivos recuperados e tempo de execução satisfatório, o *Foremost* e *MagicRescue* não foram bem-sucedidos no critério de taxas de falsos positivos, superior à 65

O *Scalpel* demonstrou ser ineficaz devido o grande aumento do quantitativo de arquivos recuperados que não condiziam aos presentes na mídia. Importante salientar que o experimento foi interrompido quando o processo atingiu 11,6% de execução, uma vez que o mesmo atingiu patamares que inviabilizam sua utilização.

4.3 Estudo de Caso: Cenário 3

O experimento, denominado Cenário 3, consistiu na simulação de recuperação de dados de tipo específico escolhidos pelo perito no momento da investigação, ou seja, caso a investigação fosse voltada apenas para descobrir ilícitos que estivessem registrados em documentos eletrônicos. Neste cenário as configurações dos *softwares* de recuperação de dados formatados foram ajustadas para recuperar documentos.

A Tabela 5 expõe os resultados apresentados no Cenário 3. Vale ressaltar que alguns *softwares* definem documentos não apenas arquivos de texto (.doc, .odt, .txt, etc), mas também, planilhas eletrônicas, documentos de apresentação, etc. Portanto, nesta avaliação, não é de suma importância a quantidade de arquivos recuperados, uma vez que a base é extensa (16.000 arquivos) e detém diversas extensões de arquivos. O foco principal é verificar a quantidade de falsos positivos.

Nesta análise, os *softwares* que apresentaram maior eficácia na recuperação de arquivos formatados foram *Recuva* e *Autopsy*, respectivamente. Ambos apresentaram resultados satisfatórios acima de 99% de arquivos recuperados corretamente. No entanto, um agravante que rebaixa os resultados do *Autopsy* é seu tempo de execução prolongado, uma vez que o mesmo realiza um processo de análise meticulosa na imagem antes de começar extração dos dados.

É notada a ausência do *PhotoRec* entre os *softwares* com melhores resultados. Pois, ao proceder com o experimento, foi identificado que o mesmo não apresentava um recurso de recuperação de arquivos específicos. Uma vez iniciado o processo de recuperação, tudo que estava na imagem ou na mídia será passível de recuperação. Por outro lado, o *Foremost*, ao especificar o interesse em recuperar arquivos da extensão .doc ou .docx, o mesmo retornou zero registros. Quando especificados arquivos do tipo .odt o software finaliza, sinalizando uma não identificação dessa extensão.

Diferente dos resultados obtidos nos cenários anteriores, o Magic Rescue apresentou um percentual razoável de verdadeiros positivos no experimento, superando seus pares anteriores nesse aspecto. Por outro lado, foi insatisfatório nos aspectos de tempo de execução e tamanho dos dados recuperado. O mesmo cenário de resultados aquém se repete com o

Table 4. Cenário 2: Resultados dos softwares periciados

Softwares	Qtd. de arquivos gerados	Tempo de Execução	Tamanho do diretório recuperado	Falsos positivos (%)	Verdadeiros positivos (%)	Verdadeiros positivos repetidos (%)
Recuva	13952	13:14:54	117 GB	1,1%	98,8%	0,05%
Foremost	20150	4:06	2,2 GB	65,4%	34,5%	0%
Scalpel	508482	13:54:15	100,9 GB	>99%	<1%	0%
Magic Rescue	15950	2:04:47	2,2 GB	80,1%	19,9%	0%
Photorec	14830	06:50	5 GB	8,2%	>91,7%	0%
Autopsy	32149	14:20:03	193 GB	10%	90%	44,5%

Scalpel, uma vez que não entrega um bom percentual de verdadeiros positivos, além de manter o maior tempo de execução e tamanho dos arquivos.

4.4 Verdadeiros Positivos

Já citado anteriormente, um dos principais agravantes na abordagem de data carving é o acréscimo de arquivos que não existiam e foram recuperados. Trata-se de indevida localização do cabeçalho e rodapé de algum arquivo da mídia periciada (falsos positivos). Para responder à pergunta da pesquisa, a Figura 4 ilustra os percentuais de verdadeiros positivos encontrados utilizando diversos *softwares* nos três cenários propostos.

Outro aspecto importante a ser considerado é a quantidade de arquivos recuperados em relação ao número de arquivos existentes em cada cenário. As imagens .dd que foram periciados continham 1.000 e 16.000 arquivos para os cenários 1 e 2, respectivamente. A Figura 5 ilustra esta informação. Considerando que os valores que estão à esquerda (negativos) indicam que os *softwares* recuperaram menos que o esperado. Já os valores à direita (positivos) representam os *softwares* recuperaram arquivos além do quantitativo devido.

Na interpretação desenvolvida pelo presente trabalho, falsos positivos remetem a arquivos recuperados erroneamente, corrompidos ou que não existiam anteriormente na mídia original. No caso de arquivos duplicados, trata-se de arquivos verdadeiros, porém o *software* recuperador não é dotado de métodos para eliminá-los. Tanto falsos positivos quanto arquivos duplicados diminuem a eficiência da solução, pois exigem recursos computacionais em excesso. Por recursos computacionais, denotam-se espaço de armazenamento de dados e tempo de execução.

Porém, os falsos positivos apresentam um potencial problemático adicional. Caso o perito forense digital não os elimine artesanalmente, todo o trabalho estará comprometido. A investigação poderia se basear em arquivos que nunca existiram. A presente interpretação, portanto, estabelece como fundamental a segregação entre arquivos duplicados e aqueles falsos positivos.

Para fins de comparação, nesta análise, o ideal são os *softwares* que apresentam resultados em zero ou próximo. Isso significa que recuperaram a quantidade ideal de arquivos, nem a menos, nem além do que existia na mídia periciada. Neste aspecto, o *PhotoRec* e *Recuva* apresentaram melhores resultados.

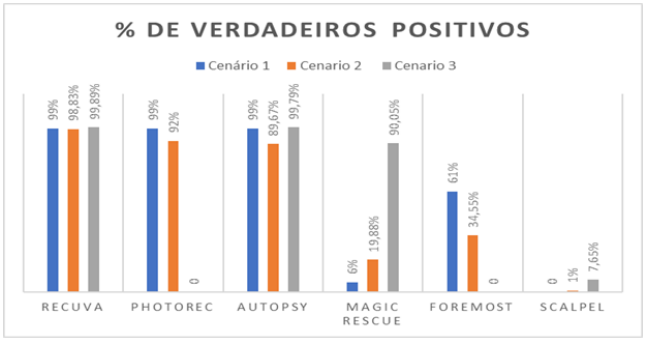

Figure 4. Percentual de Verdadeiros Positivos. Fonte: Autoria própria

Por fim, não consta na Figura 4 os resultados do Scalpel, uma vez que o resultado do mesmo destoa de todos os outros. Também não foi inserido os resultados do Cenário 3, visto que, como explicado anteriormente, não há um padrão entre os softwares sobre o que é considerado arquivos do tipo documentos, não sendo possível ter uma quantidade mínima de arquivos a serem recuperados.

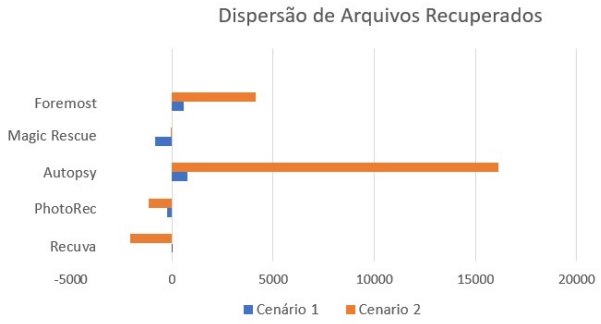

Figure 5. Dispersão de Arquivos Recuperados. Fonte: Autoria própria

5. CONCLUSÃO E TRABALHOS FUTUROS

Foram analisados diversos *softwares* de *data carving* que são amplamente encontrados na literatura em cenários mais simples e complexos. Dadas as métricas extraídas, espera-

Table 5. Cenário 3: Resultados dos softwares periciados

Softwares	Qtd. de arquivos gerados	Tempo de Execução	Tamanho do diretório recuperado	Falsos positivos (%)	Verdadeiros positivos (%)	Verdadeiros positivos repetidos (%)
Recuva	1969	09:43	487 MB	<1%	>99%	0%
Foremost	0	1:04	0	-	-	-
Scalpel	6862	04:27:07	34,7 GB	92,3%	7,6%	36,5%
Magic Rescue	3148	51:06	1 GB	9,9%	90,1%	0%
Photorec	-	-	-	-	-	-
Autopsy	3915	14:20:19	691,9 MB	<1%	>99%	1,5%

se que o trabalho proporcione ao perito a orientação sobre quais *softwares* utilizar, considerando um contexto específico. Deixando-o ciente das implicações potenciais e o percentual de falsos positivos que possam ser gerados.

O acréscimo de arquivos que não existiam e foram recuperados, resultante da indevida localização seja do cabeçalho ou do rodapé de alguma mídia periciada (falsos positivos), é um dos principais agravantes na abordagem de data carving. Para responder à pergunta motivadora dessa pesquisa, o presente trabalho apresentou percentuais de falsos e verdadeiros positivos encontrados utilizando diversos softwares nos três cenários propostos.

Outro aspecto importante é a quantidade de arquivos que foram recuperados em relação à quantidade de arquivos existentes em cada cenário. As partições formatadas propositalmente continham 1.000 e 16.000 arquivos para os Cenários 1 e 2, respectivamente. Para fins de comparação nesta análise, o ideal são os *softwares* que apresentam resultados em zero ou próximo. A constatação desses resultados significa a recuperação da quantidade ideal de arquivos que existia na mídia periciada. Neste aspecto, o *PhotoRec* e *Recuva* apresentaram melhores resultados.

O presente estudo avaliou a eficiência tanto de *softwares* de código aberto como *softwares* proprietários sem disponibilidade de código-fonte. Enfatiza-se que os motivos de diferenças de desempenho entre os *softwares* implica em conhecer as particularidades de cada um. Como algumas ferramentas são código-fonte fechado, torna-se inviável fazer conjecturas quanto às razões de suas limitações.

Quanto aos *softwares* de código aberto, tal qual *Scalpel* e *Foremost*, é possível fazer conjecturas quanto ao seu desempenho. Por exemplo, no próprio arquivo de configuração do *Scalpel*, há um alerta de que a tentativa de recuperação de alguns tipos de arquivos pode gerar uma enorme quantidade de falsos positivos. Por consequência, um volume discrepante de falsos positivos podem acarretar uma lentidão até que o aplicativo conclua seu trabalho.

Em termos didáticos, a quantidade de arquivos falsos positivos gerados está atrelada a um tempo de execução demasiado. Nessas situações de falsos positivos em demasia, a perícia forense digital se torna complexa devido ao consumo excessivo de recurso computacional em relação a armazenamento de dados gerados e a tempo de execução. Na sequência, pode haver um grande atraso na conclusão do trabalho porque caberá ao perito forense digital distinguir os arquivos existentes daqueles gerados erroneamente.

Quando necessário a utilização de técnicas de *data carving* para coletar evidências de um processo judicial, é de fundamental importância que o perito digital disponha recursos ferramentais que propicie maior eficácia ao seu trabalho. Considerando que o perito tende a trabalhar com grandes volumes de dados oriundos de mídias e dispositivos que são investigados. O tempo gasto nas etapas de análises, extração e recuperação de dados é crucial para entrega de respostas em tempo hábil à justiça.

Como proposta de continuidade da pesquisa apresentada, os autores pretendem: (i) Ampliar o quantitativo de *softwares* analisados, bem como explorar outros cenários de teste. (ii) Expandir a base de dados, utilizada como suporte para a análise de mídias periciadas, incluindo outras extensões de arquivos. (iii) Coletar métricas adicionais relacionadas à facilidade de utilização dos softwares de data carving. (iv) Experimentar técnicas de classificação com Machine Learning, redes neurais e redes extremas, como contemplado em [25]. (v) Ampliar a análise de desempenho dos *softwares* para outros formatos de arquivos, tipicamente utilizados para uso pessoal, tais como planilhas eletrônicas e de apresentação.

Contribuições dos autores

S.M.L.L. e S.M.M.F. conceberam a proposta de avaliação de ferramentas para recuperação de dados formatados. R.K.P.S. e I.A.B. construíram o conjunto de dados, planejaram os cenários de teste e o arcabouço metodológico. R.K.P.S. executou os experimentos, analisou os resultados alcançados e escreveu o manuscrito com apoio de S.M.L.L. e S.M.M.F. Todos os autores discutiram os resultados e contribuíram para o manuscrito final e para a interpretação dos resultados. Todos os autores forneceram feedback crítico e ajudaram a moldar a investigação, a análise e o manuscrito.

References

[1] STANKOVIć, M.; KHAN, T. Digital forensics tool evaluation on deleted files. digital forensics and cyber crime. *ICDF2C Springer*, v. 508, 2022.

[2] BLASKOVIC, A. K. et al. Cybercrime and intellectual property theft: An analysis of modern digital forensics. *Proceedings of the Future Technologies Conference (FTC), Springer International Publishing*, v. 2, 2023.

[3] POVAR, D.; BHADRAN, V. K. Forensic data carving. In: BAGGILI, I. (Ed.). *Digital Forensics and Cyber Crime*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 137–148. ISBN 978-3-642-19513-6.

[4] HANIS, F. M. et al. A language-independent approach to classification of textual file fragments: Case study of persian, english, and chinese languages. In: *2021 11th International Conference on Computer Engineering and Knowledge (ICCKE)*. [S.l.: s.n.], 2021. p. 254–259.

[5] GLAYDYSHEV, P.; JAMES, J. I. Decision-theoretic file carving. *Digital Investigation - Science Direct*, v. 22, 2017.

[6] WEI, Y.; ZHEN, N.; XU., M. An automatic carving method for rar file based on content and structure. *Second International Conference on Information Technology and Computer Science*, 2010.

[7] LAURENSON, T. Performance analysis of file carving tools. *IFIP Advances in Information and Communication Technology*, Springer New York LLC, v. 405, p. 419 – 434, 2013.

[8] SARI, S.; MOHAMAD, K. A review of graph theoretic and weightage techniques in file carving. *Journal of Physics: Conference Series*, Institute of Physics Publishing, v. 1529, 2020.

[9] SILBERSCHATZ, A.; GALVIN, P. B.; GAGNE, G. *Operating system concepts*. 10. ed. [S.l.]: Wiley, 2018.

[10] RAVI, A.; T., R. K.; MATHEW, A. R. A method for carving fragmented document and image files. *International Conference on Advances in Human Machine Interaction, India*, 2016.

[11] TANG, Y. et al. Recovery of heavily fragmented jpeg files. *Digital Investigation*, Elsevier Ltd, v. 18, p. S108 – S117, 2016.

[12] LAURENSON., T. Performance analysis of file carving tools. *Security and Privacy Protection in Information Processing Systems. Springer*, v. 405, 2013.

[13] NURHAYATI; FIKRI, N. The analysis of file carving process using photorec and foremost. In: *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*. [S.l.: s.n.], 2017. p. 1–6.

[14] KARRESAND, M.; DYRKOLBOTN, G. O.; AXELSSON., S. An empirical study of the ntfs cluster allocation behavior over time. *Forensic Science International: Digital Investigation*, v. 33, 2020.

[15] WU, T.; BREITINGER, S. O. F. Digital forensic tools: Recent advances and enhancing the status quo,. *Forensic Science International: Digital Investigation*, v. 34, 2020.

[16] UZUN, E.; SENCAR, H. T. Jpg scraper : An advanced carver for jpeg files. *IEEE Transactions on Information Forensics and Security*, v. 15, 2020.

[17] HILGERT, J.; LAMBERTZ, M.; RYBALKA, R. S. M. Syntactical carving of pngs and automated generation of reproducible datasets. *Digital Investigation*, v. 29, 2019.

[18] RAVI, A.; KUMAR, T. R.; MATHEW, A. R. A method for carving fragmented document and image files. In: *2016 International Conference on Advances in Human Machine Interaction (HMI)*. [S.l.: s.n.], 2016. p. 1–6.

[19] PEREIRA, E. et al. Análise de métodos para o tratamento de arquivos falso-positivos a partir de ferramentas de recuperação de dados digitais: Uma revisão sistemática da literatura. p. 1–10, 2019.

[20] PRODANOV, C. C.; FREITAS, E. C. D. *Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição*. [S.l.]: Editora Feevale, 2013.

[21] ALI, K. M. Digital forensics best practices and managerial implications. In: *2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks*. [S.l.: s.n.], 2012. p. 196–199.

[22] SCHNEIDER, J.; EICHHORN, M.; FREILING, F. Ambiguous file system partitions. *Forensic Science International: Digital Investigation*, v. 42, 2022.

[23] 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.

[24] PALMIERI, G.; ZARGARI., S. Using open source forensic carving tools on split dd and ewf files. *IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2017.

[25] LIMA, S. M. L. et al. Next-generation antivirus endowed with web-server sandbox applied to audit fileless attack. *Springer Nature*, 2022.