Revista ARACÊ
ISSN: 2358-2472

DEJAVU FORENSICS: APRIMORANDO A RECUPERAÇÃO DE DADOS JPEG E PNG FORMATADOS USANDO MÁQUINAS DE VETOR DE SUPORTE

https://doi.org/10.56238/arev7n3-301

Data de submissão: 28/02/2025 **Data de publicação:** 28/03/2025

Islan Amorim Bezerra
Doutoranda em Engenharia da Computação
Universidade de Pernambuco – UPE
E-mail: iab@ecomp.poli.br
ORCID: https://orcid.org/0000-0002-8920-6691
Lattes: http://lattes.cnpq.br/0162367027095181
Rubens Karman Paula da Silva
Doutoranda em Engenharia da Computação
Universidade de Pernambuco – UPE
E-mail: rkps@ecomp.poli.br
ORCID: https://orcid.org/0000- 0001-9388-9889
Lattes: http://lattes.cnpq.br/6372811203248481
Sidney Marlon Lopes de Lima
Pós-Doutorado em Engenharia da Computação
Universidade de Pernambuco – UPE
E-mail: smll@ecomp.poli.br
ORCID: https://orcid.org/0000-0002- 4350-9689
Lattes: http://lattes.cnpq.br/0323190806293435
Sergio Murilo Maciel Fernandes
Doutoramento em Engenharia Informática
Universidade de Pernambuco – UPE
E-mail: smurilo@ecomp.poli.br
ORCID: https://orcid.org/0000-0002-5922-4119
Lattes: http://lattes.cnpq.br/4520293519781462
Carolina de Lira Matos
Graduando em Engenharia Elétrica e Eletrônica
Universidade Federal de Pernambuco – UFPE
E-mail: carolina.liram@ufpe.br
ORCID: https://orcid.org/ 0009-0009-2069-869X
jheklos Gomes da Silva
Doutoranda em Engenharia da Computação
Universidade de Pernambuco – UPE
E-mail: jheklos.gomess@upe.br
ORCID: https://orcid.org/0000-0002-7741-8849
Lattes: http://lattes.cnpq.br/3048498549751735

RESUMO
Com os avanços tecnológicos, os crimes virtuais estão ocorrendo com mais frequência. Quando um equipamento digital é roubado, perdido ou descartado, os dados permanecem armazenados nos discos, possibilitando sua recuperação. Este trabalho tem como foco a recuperação de arquivos formatados, investigando a aplicabilidade das ferramentas Foremost, Scalpel e Magic Rescue no Linux, além de uma ferramenta interna equipada com aprendizado de máquina. O objetivo é desenvolver uma ferramenta para a recuperação e validação de arquivos formatados, contribuindo para investigações de crimes digitais e trazendo novos insights sobre os métodos de recuperação. Usando o reconhecimento de padrões, o cluster é usado como entrada, atuando como um neurônio na máquina de aprendizado. O trabalho aplica aprendizado de máquina para reconhecer padrões em blocos/clusters. No cenário

Revista ARACÊ
ISSN: 2358-2472

"simples", a classificação é binária (classe vs. contraclasse), metodologia desenvolvida por Pavel (2017). No cenário "complexo", foi utilizado o método um-contra-todos, com um banco de dados de 16.000 arquivos. A pesquisa apresenta uma abordagem que combina aprendizado de máquina e ciência de dados para recuperar dados formatados. A ferramenta interna atinge uma taxa de recuperação de mais de 96% para arquivos PNG e JPEG formatados, rodando em segundos.

Palavras-chave: Escultura de dados. Forense digital. Cibercrime. Recuperação de dados. Aprendizado de máquina.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

Na era digital, o acesso à informação está impulsionando avanços significativos em muitas áreas do conhecimento. Esses avanços contribuem para o desenvolvimento de ferramentas para otimizar processos. Por outro lado, a ascensão e a constante evolução da tecnologia criaram um ambiente propício ao surgimento e disseminação do crime digital. Esses casos são caracterizados por atividades ilegais que comprometem a integridade, confidencialidade e segurança dos usuários.

A perícia digital é responsável pelo esclarecimento de crimes digitais com base na busca de provas e fatos que incriminam os suspeitos e protegem as vítimas. A busca por evidências pode ser realizada em dispositivos como smartphones e redes de computadores. O foco está na aplicação da lei, para que a investigação seja conduzida adequadamente e constitua provas admissíveis em um tribunal.

A relação entre crime digital e recuperação de dados é dada pelo uso do usuário. Os usuários de hoje usam seus dispositivos, como smartphones e computadores, como uma espécie de arquivo digital. Vários programas e aplicativos são baixados para uma variedade de funções. Os arquivos são salvos para uso posterior e aqueles que não devem ser salvos devem ser excluídos.

A digitalização de metadados é vital em investigações de crimes digitais. Ele fornece informações detalhadas sobre arquivos, listando suas datas de criação, modificação e acesso, tipo de dispositivo e permissões. Esses dados ajudam os especialistas forenses a investigar atividades suspeitas, como acesso não autorizado e adulteração de documentos, e também os ajudam a reconstruir cronogramas. No entanto, cibercriminosos sofisticados podem dificultar essa análise excluindo ou substituindo metadados por meio da formatação do dispositivo. Isso dificulta a recuperação de informações críticas e complica o trabalho dos investigadores no rastreamento de evidências digitais.

A escultura de arquivos é uma técnica de recuperação de dados. É uma ferramenta em perícia digital e serve como opção final quando outros métodos de recuperação não estão mais disponíveis. Essa abordagem torna-se essencial nos casos em que é impossível reconstruir metadados do sistema de arquivos, como com discos rígidos formatados.

A escultura de arquivos extrai fragmentos de arquivos sem usar estruturas do sistema de arquivos. Esse processo analisa os dados brutos no disco, identificando assinaturas ou padrões exclusivos de diferentes tipos de arquivos. Esse recurso permite que os investigadores recuperem arquivos de uma unidade formatada.

Quando usada incorretamente, a escultura de arquivos pode se tornar uma ferramenta poderosa nas mãos de cibercriminosos. Se um dispositivo eletrônico for perdido ou roubado, os dados geralmente permanecem no armazenamento. Agentes mal-intencionados podem recuperá-lo mesmo após a exclusão ou formatação. Fragmentos desses arquivos podem ser recuperados, o que pode expor

Revista ARACÊ
ISSN: 2358-2472

dados confidenciais. O principal risco está nas informações recuperadas. Pode ocorrer uso indevido de arquivos confidenciais restaurados, imagens pessoais ou dados bancários, facilitando fraudes e outros crimes.

O objetivo principal deste artigo é explorar a aplicação de técnicas de escultura de arquivos em perícia digital, com foco na recuperação de dados formatados em cenários envolvendo crimes digitais. Ao analisar a eficácia de ferramentas como Foremost, Scalpel e Magic Rescue, bem como apresentar uma ferramenta inovadora baseada em aprendizado de máquina, Dejavu Forensics, este estudo visa aprimorar o processo de recuperação de arquivos formatados, particularmente os formatos JPEG e PNG. Além disso, a pesquisa busca demonstrar como o aprendizado de máquina, especificamente as máquinas de vetores de suporte (SVM), pode melhorar a precisão e a eficiência da recuperação de dados. Em última análise, este trabalho contribui para o campo da perícia digital, fornecendo insights sobre os desafios de recuperar dados formatados e oferecendo soluções que podem auxiliar na investigação de crimes digitais.

2 REFERENCIAL TEÓRICO

2.1 A RELAÇÃO ENTRE CRIME DIGITAL E DADOS FORMATADOS

O crime digital inclui vários atos maliciosos que comprometem sistemas de computador, dados e operações online. O aumento dos ataques cibernéticos mostra que devemos entender seu impacto no crime digital e na recuperação de dados. O crime digital é uma questão complexa que inclui a exploração de falhas de segurança, o uso de malware e a realização de ataques em um ambiente digital. Também envolve ações que comprometem os dados do usuário. O aumento dos ataques digitais demonstra a evolução das táticas dos criminosos em espaços online.

Outra situação crítica com altas taxas de incidência é a sextorsão. De acordo com o Observatório de Crimes Cibernéticos, a extorsão sexual refere-se à extorsão sexual, como vídeos ou fotos íntimas. Os criminosos obtêm acesso ao conteúdo íntimo de suas vítimas e as coagem com a ameaça de exposição. Na maioria dos casos, parceiros ou ex-parceiros cometem esse crime, mas também pode ser feito por hackers com quem a vítima compartilhou conteúdo erótico online (Cybercrime Observatory, 2020). Este crime destaca a negligência da frase "olhe e depois exclua". Mesmo que o arquivo seja excluído, os dados permanecem no dispositivo e podem ser recuperados posteriormente. Nesses casos, as ferramentas forenses digitais, como o Dejavu Forensics, desempenham um papel vital, permitindo a recuperação de dados formatados e, ao mesmo tempo, garantindo a adesão aos padrões forenses. Isso garante que os dados recuperados possam ser usados como prova em investigações legais, mantendo a integridade e a cadeia de custódia exigidas nos

Revista ARACÊ
ISSN: 2358-2472

procedimentos forenses.

A recuperação de dados formatados falhará se o usuário substituir o conteúdo da partição pouco a pouco, esterilizando os dados. Esse processo torna impraticável recuperar os dados originais e os investigadores não podem extrair nenhuma informação relevante para a investigação.

Como estudo de caso, dois estudantes de pós-graduação do Laboratório de Ciência da Computação do Instituto de Tecnologia de Massachusetts (MIT) publicaram um relatório sobre um experimento em que compraram 158 discos rígidos usados, dos quais apenas 129 funcionaram. Destes, apenas 12 foram devidamente "esterilizados". De resto, foi possível recuperar mais de 5.000 números de cartão de crédito, pastas de e-mail, transações bancárias, endereços de e-mail e registros hospitalares (Garfinkel, 2003).

No entanto, conforme ilustrado no experimento conduzido pelos dois alunos do MIT, muitos usuários, intencionalmente ou não, não aplicam a esterilização de dados corretamente. A investigação encontrou milhares de detalhes pessoais confidenciais em discos rígidos supostamente excluídos (Garfinkel, 2003).

Existem várias ferramentas para esterilizar dados, mas elas não são nativas dos sistemas operacionais Windows ou smartphones. No Android, os usuários podem realizar uma formatação leve pressionando dois botões por 10 segundos, o que é fácil. Mas a esterilização é mais complexa, pois não há metodologias nativas disponíveis. Normalmente, os usuários precisam conectar o smartphone a um computador desktop, onde os comandos de linha devem ser usados para esterilização. Além disso, esse procedimento tem desvantagens, pois pode anular a garantia e o Android impede o acesso a sistemas operacionais externos.

Nesse sentido, os dados formatados, por sua natureza estruturada e crítica, são um dos elos mais visados nas atividades ilícitas no ambiente digital por meio de diversas técnicas de invasão e exploração. A possibilidade de recuperação de dados formatados torna-se um alvo por se tratar de um ambiente rico em informações sensíveis, e o acesso ou manipulação inadequada pode ter consequências negativas.

Portanto, a investigação de crimes digitais passa a ser de responsabilidade da computação forense, que combina elementos do direito e da ciência da computação para obter rastros e evidências de crimes digitais (Casey, 2011). A perícia digital é uma atividade recente, com suas primeiras práticas surgindo na década de 1980, quando surgiram os primeiros casos de vírus em redes de comunicação. Nos anos seguintes, as investigações evoluíram para casos de pedofilia e, posteriormente, para crimes cibernéticos mais amplos (Hannan, 2004).

Na análise forense, a recuperação de dados é o processo de restauração de arquivos perdidos,

Revista ARACÊ
ISSN: 2358-2472

o que ajuda os profissionais de informática a obter novos insights sobre a análise de métodos de recuperação de arquivos formatados. Este método é definido como o processo de recuperação de dados excluídos ou formatados logicamente (Vacca, 2006).

2.2 VISÃO GERAL DA RECUPERAÇÃO DE DADOS FORMATADOS

A recuperação de dados formatados é possível. Quando um usuário exclui um arquivo, ele é logicamente excluído do sistema de armazenamento. O sistema operacional (SO) sabe quais partes do dispositivo o arquivo ocupa. Em termos simples, o sistema operacional apenas modifica o espaço usado por esse arquivo, tornando-o disponível para uso. No entanto, os dados permanecem no dispositivo. Portanto, é possível recuperar os dados (Pal, 2009).

A tabela de alocação é atualizada para indicar que os blocos atribuídos anteriormente ao arquivo removido estão livres. No entanto, os dados formatados logicamente permanecem no dispositivo. Os dados permanecem até, por exemplo, um novo arquivo substituí-los ou quando um arquivo existente é expandido.

Mesmo que um usuário exclua os dados usando o comando delete, os dados podem permanecer no dispositivo. O mesmo acontece com o comando format, que é usado para preparar logicamente o disco para uso, mas não garante que os dados serão destruídos (James, 2006).

As técnicas forenses digitais usam informações residuais como recurso para recuperar arquivos, como fragmentos que o usuário pensou terem sido excluídos. Como tal, a recuperação de dados pode atuar como evidência potencial de crime digital. Deve-se enfatizar que essas técnicas são usadas para fins positivos e negativos.

2.3 ESCULTURA DE DADOS E SUAS LIMITAÇÕES

Escultura é um termo geral para extrair arquivos de dados brutos de um sistema de arquivos. Baseia-se nas características específicas do formato do arquivo (Alherbawi, 2013). A escultura de dados, na computação, refere-se à recuperação de arquivos excluídos. No entanto, corresponde a uma técnica realizada pela localização de assinaturas conhecidas.

Quando um disco é formatado logicamente, é criada uma tabela que atua como uma espécie de mapa. Este mapa guia os cabeçotes de leitura e gravação para as posições corretas onde cada arquivo gravado está localizado. Esta tabela contém o sistema de arquivos, que determina: (i) como os arquivos serão acessados (setores) e (ii) o tamanho dos clusters.

O armazenamento do computador é organizado em unidades chamadas setores. O sistema de arquivos, por sua vez, agrupa esses setores em unidades menores, conhecidas como blocos/clusters

Revista ARACÊ
ISSN: 2358-2472

(Ali, 2018). Os clusters são caracterizados por terem cabeçalhos que funcionam como um marcador para o início dos arquivos. Essa assinatura descreve os métodos do sistema de arquivos, bem como o tipo de arquivo e o conteúdo.

Essas assinaturas consistem em um cabeçalho e rodapé hexadecimais, e cada tipo de arquivo tem um identificador exclusivo. Um cabeçalho identifica os bytes no início do arquivo, enquanto o rodapé identifica os bytes no final do arquivo.

Em resumo, para cada tipo de arquivo ou categoria de arquivos, é necessária uma técnica diferente. Por meio do carving de dados, é necessário verificar a estrutura dos clusters para decidir se eles são consistentes e se ainda podem ser considerados uma unidade coerente do sistema de arquivos (Sencar, 2009). Carrier (2005) descreve um sistema de arquivos como dados estruturais e organização do usuário de forma que a máquina possa encontrá-lo.

As técnicas de escultura de dados são usadas com mais frequência para recuperar arquivos que não estão alocados na unidade. O espaço alocado, nesse sentido, refere-se a uma partição na unidade. Esta partição não mostra nenhuma informação de arquivo, como se estivessem danificadas ou ausentes (Pal, 2009). A abordagem de escultura de dados foi iniciada pelo Laboratório Forense de Computadores de Defesa (DCFL) produzindo "Carv This". Em seguida, Kriss Kendall e Jesse Kornblumt apresentaram 'Foremost', que corresponde a uma ferramenta de código (Sari, 2020).

De acordo com Kent, o processo forense compreende quatro fases, conforme descrito na Figura 1 (Kent, 2006). A fase de coleta tem como objetivo rotular, identificar e adquirir os dados. A fase de exame envolve a automação de processos e a combinação de dados para extrair os dados relevantes. A análise usa métodos e técnicas para adquirir informações com base nos dados coletados. Por fim, os resultados obtidos correspondem aos relatórios finais da análise, que incluem os procedimentos utilizados e possíveis melhorias que podem ser feitas.

Figura 1 - Diagrama do processo forense

Fonte: Elaborado pelos próprios autores

Revista ARACÊ
ISSN: 2358-2472

Uma das principais características dessa técnica é a assinatura da estrutura do arquivo. Os arquivos de extensão PDF, por exemplo, têm a assinatura inicial, o cabeçalho. Ou seja, os arquivos em formato PDF sempre começam da mesma maneira, com o mesmo cabeçalho. Os arquivos JPEG, por sua vez, possuem em sua estrutura o cabeçalho hexadecimal "0xFFD8" e o rodapé "0xFFD9", conforme mostrado na Figura 2. Isso permite distingui-lo de outros tipos de arquivos, com base em um exame do conteúdo. A pesquisa de sequência de cabeçalho e rodapé também é conhecida como "números mágicos".

Os sistemas de arquivos são responsáveis pelo gerenciamento da estrutura. Além disso, eles alocam clusters sequenciais ou não. A ausência de uma cadeia de caracteres nos clusters indica um problema de armazenamento conhecido como fragmentação. Esses fragmentos podem ser distribuídos em qualquer ordem, dependendo do sistema, dos tamanhos dos arquivos e do tamanho do cluster.

Apesar dos benefícios que o Data Carving pode oferecer, existem algumas limitações significativas. Tais limitações podem interferir na garantia da eficácia da recuperação de dados formatados. Sendo: i. a geração de arquivos falso-positivos, que geralmente são arquivos corrompidos; ii. a possibilidade de descartar arquivos que, de fato, existem no sistema; iii. além de desconsiderar possíveis fragmentos.

2.3.1 Limitações de Escultura de Dados: Falsos Positivos

Durante um processo de investigação criminal é necessário considerar o uso de métodos e protocolos. O objetivo é garantir a eficácia dos dados que estão sendo recuperados. É necessário garantir a eficácia durante este processo. É importante que a ferramenta utilizada gere o menor número possível de arquivos falso-positivos.

Quando um processo de recuperação de dados digitais é executado, arquivos reais são gerados. No entanto, arquivos incorretos ou corrompidos podem ser gerados, chamados de falsos positivos (Laurenson, 2013). Falsos positivos são arquivos que nunca existiram e que podem ser gerados em grande número.

Por não existirem, esses arquivos falsos podem ser muito grandes. Sua presença tende a ser sinônimo de dificuldade no processo realizado pelo perito criminal.

Quanto maior o número de falsos-positivos, maior a dificuldade em separar arquivos reais daqueles que não existem.

Uma das razões para o grande número de falsos positivos são as coincidências entre os cabeçalhos do cluster. Por exemplo, considere um determinado cluster contendo os caracteres iniciais

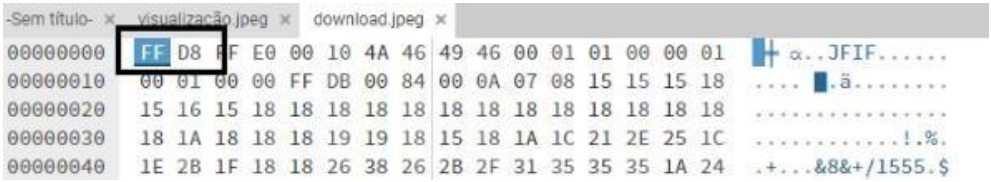
Revista ARACÊ
ISSN: 2358-2472

"\%PDF". Neste cluster, haveria recuperação de um arquivo pdf excluído anteriormente. Mas pode ser uma simples coincidência que haja uma cadeia de caracteres igual ao cabeçalho no início do cluster de dispositivos de memória sondados.

O Data Carving começaria a recuperar um arquivo pdf que, na verdade, nunca existiu. Como o arquivo não existe, pode ser que seu rodapé não esteja presente em nenhum dos clusters a seguir. Como consequência, o arquivo falso pode ter tamanhos gigantescos, por exemplo, o arquivo falso pode ter o tamanho da partição.

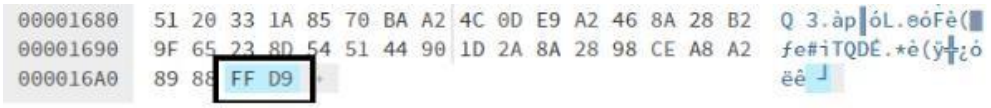
A Escultura de Dados determina os critérios de parada ao pesquisar o rodapé nos arquivos de configuração. Atua como uma medida paliativa. A ferramenta Bisturi, baseada em Data Carving, a partir de sua configuração elimina metadados indesejados na primeira fase. Ele avisa que existem alguns cabeçalhos e rodapés que podem gerar muitos arquivos falsos positivos. Além de determinar os tamanhos mínimo e máximo que os arquivos podem conter.

Figura 2 - Escultura de dados: cabeçalho do arquivo jpeg



Fonte: Elaborado pelos autores

Figura 3 - Escultura de dados: Rodapé do arquivo jpeg.



Fonte: Elaborado pelos autores.

2.3.2 Limitações de entalhe de dados: arquivos existentes ignorados

Os sistemas de arquivos têm metadados que descrevem suas respectivas estruturas. Cada estrutura tem uma sequência específica em cabeçalhos e rodapés. Isso determina o ID do arquivo.

Durante a identificação da assinatura do cabeçalho, os arquivos podem ser descartados. O Data Carving inicia o processo de recuperação. Este processo só é encerrado ao encontrar o rodapé do arquivo. No entanto, o cabeçalho acaba sendo ignorado. Como o Data Carving está procurando o final do arquivo correspondente; o rodapé.

Nesse contexto, o DECA – Decision-Theoretic Carving – surge como uma ferramenta capaz de reduzir o número de falsos positivos (Gladyshev, 2017). Isso acontece por meio de técnicas de aprendizado de máquina - especificamente SMV. A primeira etapa consiste em buscar o cabeçalho no

Revista ARACÊ
ISSN: 2358-2472

início do cluster. Na segunda etapa, o aprendizado de máquina é executado. Isso ocorre entre o cluster analisado e seu respectivo tipo de arquivo por meio do uso do reconhecimento de padrões.

2.3.3 Impacto do tamanho do bloco/cluster na recuperação de dados

O tamanho dos blocos ou clusters nos sistemas de arquivos desempenha um papel importante no processo de recuperação de dados. Clusters menores permitem uma recuperação de dados mais detalhada, pois cada bloco contém uma quantidade menor de dados. No entanto, isso pode aumentar o tempo necessário para a recuperação, pois mais clusters precisam ser processados.

Por outro lado, clusters maiores aceleram o processo de leitura e recuperação, mas podem aumentar a possibilidade de perda de dados, uma vez que mais informações são armazenadas em cada bloco. Isso pode afetar a precisão da recuperação, especialmente quando há corrupção de dados em um bloco específico.

Estudos sugerem que o tamanho do cluster afeta as ferramentas de recuperação de dados. Isso afeta o tempo e a precisão do processo (Alherbawi, 2013). As ferramentas que lidam com clusters maiores tendem a ser mais rápidas, mas podem enfrentar desafios em termos de integridade dos dados recuperados.

Em nossa estrutura, a ferramenta Sleuth kit é usada para recuperar clusters e seus parâmetros associados, como tamanho e localização. Escolhemos a ferramenta do kit Sleuth por sua robustez e eficiência na análise forense. Também veremos sua integração com outras ferramentas forenses usadas no estudo.

2.3.4 Limitações de entalhe de dados: fragmentos de cluster com

Do ponto de vista forense, é importante estar ciente dos arquivos fragmentados. Os arquivos podem exibir fragmentação quando um ou mais clusters não fazem parte do arquivo verificado. A fragmentação geralmente ocorre quando os dados são armazenados parcialmente. O que pode acontecer devido à redução do espaço livre ou até mesmo à exclusão dos arquivos.

Um ponto de fragmentação é o último cluster em um fragmento, ou seja, onde ocorre a fragmentação. Um segmento possui um ou mais fragmentos consecutivos que pertencem a vários arquivos (Sari,2020).

A fragmentação ocorre quando um determinado arquivo não é armazenado na sequência correta em clusters consecutivos. Ou seja, sequenciar os clusters do início ao fim resulta em recuperação incorreta de arquivos. É como se fornecesse parte, ou fragmentação, do que havia sido armazenado (Pal; Memon, 2009).

Revista ARACÊ
ISSN: 2358-2472

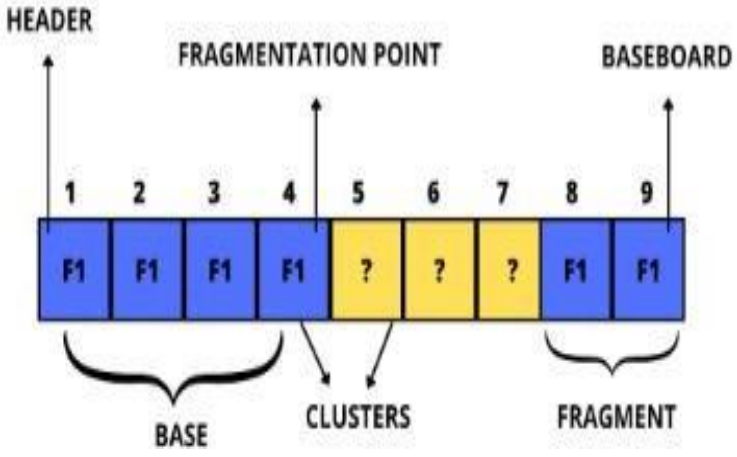
No estudo realizado por Garfinkel (2007), o autor começa identificando aproximadamente 350 discos rígidos. Os resultados indicam que as taxas de fragmentação de arquivos como e-mails, JPEG e Microsoft Word tendem a ser maiores. O que pode acontecer por vários motivos, como pouco espaço de armazenamento no disco ou até mesmo desgaste no sistema de arquivos.

Essas fragmentações também podem ser causadas por blocos defeituosos. Blocos defeituosos correspondem a áreas corrompidas no próprio disco rígido. Possivelmente isso se deve à superfície defeituosa. Nesse caso, o Data Carving pode não descartar os blocos defeituosos. Consequentemente, o arquivo recuperado será corrompido, impossibilitando a visualização e leitura do arquivo.

Além dos blocos defeituosos, ainda existem fragmentos. Eles são causados por destruições lógicas que acontecem indevidamente pelo sistema operacional. Os dispositivos de armazenamento usam algoritmos que realizam um tipo de suavização, em que o controlador realoca endereços entre blocos lógicos e físicos (Pal; Memon, 2009). Em geral, se o arquivo recuperado tiver fragmentos, possivelmente o arquivo está corrompido e não pode ser aberto.

Quando o disco tem fragmentações, há alguns pequenos grupos de clusters disponíveis para armazenamento. O que acontece é que os possíveis arquivos que ocuparão esse espaço podem ter tamanhos maiores. Como resultado, esses arquivos são armazenados e, durante o processo de armazenamento, são triturados para alocação.

Figura 4 - Data Carving: Bad Blocks..



Fonte: Adaptado de PAL, *et al.* (2009).

Além de ser causada por defeitos físicos, a corrupção de arquivos e clusters também pode ser causada por destruição lógica indevida pelo Sistema Operacional e até mesmo por mau funcionamento após seu mau funcionamento após sua atualização automática. Um exemplo disso é a não detecção de fragmentos usando a estrutura de um arquivo bitmap. Em uma estrutura de arquivo bitmap, cada parte dessa estrutura desempenha um papel importante na forma como o conteúdo é exibido para o usuário,

Revista ARACÊ
ISSN: 2358-2472

bem como as informações (metadados) geradas e usadas pelo sistema operacional para realizar sua execução (abertura).

A organização básica das informações contidas no arquivo .bmp é:

- Cabeçalho do arquivo - tamanho do arquivo, assinatura inicial (%bmp) e outros;
- Cabeçalho da imagem - altura e tamanho da largura da imagem, tipo de compactação, etc;
- Tabela de cores - quais cores são usadas pela imagem;
- Imagem (conteúdo) em pixels - o conteúdo em si;

Portanto, suponha que a ferramenta de recuperação em uso tenha encontrado uma assinatura de bitmap inicial (%bmp) e tenha iniciado a Escultura de Dados no cluster C1. Como mencionado anteriormente, a Escultura de Dados não é capaz de descartar blocos defeituosos (ou clusters com um fragmento), nem pode identificá-los. Portanto, nos deparamos com o cenário de que, em um arquivo sem fragmento, o cluster C2 seria responsável por carregar as informações do cabeçalho da imagem. Considerando que, em um cluster de arquivos fragmentado, C3 corresponderá ao cabeçalho da imagem. Portanto, quando o arquivo .bmp for aberto, ele buscará informações de altura e largura da imagem de um cluster inválido. Um dos motivos para o grande número de arquivos corrompidos é que o cluster com fragmentos contém informações necessárias para a execução do arquivo.

2.4 USANDO APRENDIZADO DE MÁQUINA PARA RECUPERAR DADOS FORMATADOS

Algoritmos de aprendizado de máquina são frequentemente usados para resolver problemas de reconhecimento de padrões. A principal característica é a capacidade de generalizar diante de dados que não foram apresentados durante o processo de treinamento.

Para exemplificar algumas capacidades da IA podemos citar: processamento de linguagem natural, representação do conhecimento, raciocínio automatizado, visão computacional e robótica. Dentro dessas capacidades, há uma ampla gama de aplicações e linhas de estudo. Um deles, de interesse neste artigo, é o aprendizado de máquina para fins de recuperação de dados formatados.

O aprendizado de máquina pode automatizar de forma inteligente muitas tarefas analisando milhares de arquivos, extraindo recursos deles e ponderando-os estatisticamente. O aprendizado de máquina pode fazer muito para melhorar a segurança dos dispositivos. Existem iniciativas, mas elas ainda estão nos estágios iniciais (Gladyshev, 2017).

Para superar as limitações do Data Carving, o estado-da-arte propõe extrair recursos do cluster de disco rígido preventivamente antes de incorporá-lo ao arquivo que está sendo recuperado. Torna-se possível investigar falsos positivos, além da presença de blocos defeituosos.

Usando a metodologia de reconhecimento de padrões, para extrair recursos de arquivos, o

Revista ARACÊ
ISSN: 2358-2472

cluster é usado como entrada. Ele atua como um atributo de entrada da máquina de aprendizado. Dado esse processo, o desempenho do Data Carving pode descartar clusters que apresentam fragmentos devido a blocos defeituosos. Ou seja, o arquivo analisado em um processo de reconhecimento de padrões pode ser comprometido por algum dano que danificou o disco rígido.

2.4.1 Reconhecimento de padrão supervisionado

Os avanços tecnológicos permitem um grande armazenamento de dados. Isso possibilita encontrar padrões, tendências e anomalias que tendem a transformar dados em informação e, consequentemente, a informação se torna a base para as classificações (Witten *et al.,* 2005). O reconhecimento de padrões tem como objetivo principal construir uma representação simplificada de um conjunto de dados por meio das características consideradas mais relevantes, o que possibilita sua partição em classes (Duda; Hart, 2006).

O método de aprendizado supervisionado tem como principal característica o conhecimento avançado das aulas que serão utilizadas na geração de padrões. A disposição, o número de observações, bem como o número de classes são conhecidos, com base na medida de similaridade dos dados em questão, quando comparados aos dados previamente rotulados.

Convencionalmente, existem dois métodos que podem ser usados para realizar a classificação por meio de aprendizado supervisionado. A primeira estima um modelo probabilístico baseado em dados existentes, no qual a estimativa é classificada usando o teorema de Bayes. O segundo método pressupõe as funções discriminantes que definem os chamados limites de decisão que serão usados na classificação, com base no conjunto de dados.

Dados sensoriais por meio de uma espécie de percepção de máquina. As amostras são agrupadas de acordo com os rótulos (classes) previamente definidos pelo especialista. Os padrões que eles reconhecem são numéricos, contidos em vetores, nos quais os dados do mundo real, sejam imagens, som, texto ou séries temporais, são traduzidos.

2.5 SUPORTE AOS PARÂMETROS DO CLASSIFICADOR DE MÁQUINA VETORIAL

"Support Vector Machine" (SVM) é um algoritmo de aprendizado de máquina supervisionado usado para classificação de problemas ou análise de regressão (Mehdi; Amir, 2020). A Máquina de Vetores de Suporte é uma fronteira que melhor separa as duas classes, usando a proximidade dos dados para construir um ou mais hiperplanos para classificar dados de alta dimensão.

A operacionalização de uma tarefa de classificação é baseada na separação do conjunto de dados em "treinamento" e "teste". Nos atributos que compõem os dados (features), há um campo chamado target, que define o rótulo da classe. O objetivo do SVM é treinar modelos que possam prever

Revista ARACÊ
ISSN: 2358-2472

o atributo alvo, levando em consideração apenas os atributos na etapa de treinamento (Mehdi; Amir, 2020). Se não houver conhecimento especializado prévio sobre um domínio, o SVM é um excelente primeiro método para testar (Stuart, 2013).

Além disso, a abordagem de classificação SVM pode favorecer a análise forense e trazer mais eficiência à investigação, pois priorizaria o esforço dos especialistas. Para qualquer variedade de casos, como criminal, civil, regulatório ou organizacional, em que são necessárias etapas de pesquisa, extração e análise de dados.

Os métodos de aprendizado de máquina usados pela Máquina de Vetores de Suporte permitem a aplicação a vetores com grandes tamanhos de recursos, consistindo em uma máquina de aprendizado estatístico que implementa os princípios de Minimização de Risco Estrutural - SRM para construir o hiperplano. Usando essa metodologia, um espaço de entrada de padrões não linearmente separáveis forma um novo espaço no qual as dimensões são linearmente separáveis.

O SVM pode ser definido por um termo de interceptação b e um vetor perpendicular ao hiperplano de decisão, w̃, conhecido como vetor peso. O conjunto de treinamento sendo T , contendo (n) dados e seus rótulos , sendo $x_i y_i$X o espaço de dados e Y = −1, +1, o hiperplano é definido na Eq. (1).

$$f(x) = w \cdot x + b = 0 \qquad (1)$$

O hiperplano definido pela equação apresentada separa o espaço vetorial dos dados em duas regiões: e . Consideramos a classificação usando a função de sinal.$\vec{w} \cdot \vec{x} + b \geq 0 \vec{w} \cdot \vec{x} + b \leq 0$

A figura abaixo ilustra como sendo um ponto no hiperplano x_1H1: $w \cdot x_1 + b = +1$ e como um ponto no hiperplano x_2H2: $w \cdot x_2 + b = -1$. Esta projeção de no vetor peso w corresponde à distância entre os hiperplanos apresentados, e pode ser calculada a partir da diferença entre os hiperplanos, como mostrado na Eq. (2).$(x_1 - x_1)$

$$\begin{array}{l} w \cdot x_1 + b = +1 \\ \underline{-w \cdot x_2 + b = -1} \\ w \cdot (x_1 - x_2) = 2 \end{array} \qquad (2)$$

Quando os lados da igualdade na equação acima são multiplicados por , obtendo-se o valor da distância entre os hiperplanos, correspondendo a duas vezes o valor da margem, conforme mostrado na Eq. (3).$\frac{1}{\|W\|}$

Revista ARACÊ
ISSN: 2358-2472

$$\frac{w\cdot(x_1-x_2)}{\|w\|}=\frac{2}{\|w\|} \quad (3)$$

O valor obtido é denominado margem geométrica quando tomado como seu máximo, o que equivale a dizer que o classificador ρ corresponde à largura máxima da faixa que pode ser desenhada para separar os vetores de suporte das duas classes. É necessário encontrar w e b onde é o máximo e para cada $.\rho/\|w\|(x_i,y_i)\ \in T, y_i(w_T x_i+b)\geq 1$

É comum encontrar situações em que o conjunto de dados não é linearmente separável. Isso pode ocorrer devido à presença de ruído ou dados discrepantes, bem como outliers. Nesses casos, não é possível obter um hiperplano que separe as classes sem que ocorram erros de classificação. A estratégia é encontrar aquele que produz a menor quantidade de erro. A SVM de margem flexível é um método que pode lidar com conjuntos de treinamento que têm dados que podem violar as restrições.

A distância de um dado ao discriminante é dada pela Eq. (4).x_i

$$\frac{|w\cdot x_i+b|}{\|w\|} \quad (4)$$

Da mesma forma, quando pode ser descrito na Eq. (5).$y_i\ \in\{-1,+1\}$

$$\frac{y_i(w\cdot x_i+b)}{|w|} \quad (5)$$

A distância exibida deve ser pelo menos o valor de ρ, conforme mostrado na Eq. (6).

$$\frac{y_i(w\cdot x_i+b)}{\|w\|}\geq \rho,\quad \forall i \quad (6)$$

Fixando ρ como , a restrição dada por mas a expressão maximizada torna-se . Isso equivale a minimizar , que define o problema de otimização conforme mostrado na Eq. (7).$\rho=\frac{1}{\|w\|}\forall(x_i,y_i)\in T,\quad y_i(w_T x_i+b)\geq 1\frac{2}{\|w\|^2}\frac{1}{2}\|w\|^2$

$$\arg\min_w\ \frac{1}{2}||w||^2 \quad (7)$$

Portanto, isso corresponde à formulação padrão do SVM dada como um problema de minimização: encontre w e b como a Eq.(8).

Revista ARACÊ
ISSN: 2358-2472

$\frac{1}{2}||w||_2$ is minimized and $(8)\forall(x_i, y_i) \in T, y_i(w \cdot x_i + b) \geq 1$

O problema de otimização obtido é quadrático sujeito a restrições lineares. Como a função objetivo é convexa e os pontos que satisfazem a restrição formam um conjunto convexo, esse problema apresenta um mínimo local único (Haykin, 2008). Isso pode ser resolvido introduzindo uma função Lagrangeana cujo objetivo é construir um problema dual no qual o multiplicador de Lagrange está associado à restrição $\alpha_i y_i(w_T x_i + b) \geq 1$

no problema primordial incluí-los na função objetivo (Manning; Raghavan; Schütze, 2008).

O problema duplo corresponde a uma descoberta como a Eq. (9).$\alpha_i, \ldots, \alpha_N$

$L(\alpha, w, b) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i(y_i(w \cdot x_i + b) - 1) = \quad (9)\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i\big(y_i(w \cdot x_i + b)\big) + \sum_{i=1}^{n} \alpha_i$

A função Lagrangeana deve ser minimizada, o que implica minimizar w e b e maximizando as variáveis . Há um ponto de sela no qual os gradientes da equação acima em relação a $\alpha_i w$ e b são zero e que, conforme mostrado na Eq. (10).$\alpha_i \geq 0$

$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i \qquad (10)\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_i \alpha_i y_i = 0$

Inserindo essas equações, temos a Eq. (11).

$C \arg\max_{\alpha} \sum_j \alpha - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \big(x_i \cdot x_j\big) \qquad (11)$

Dentro das restrições da Eq. (12).

$$\begin{cases} \sum_i \alpha_i y_i = 0 \\ \alpha_i \geq 0, \forall_i = 1, \ldots, n \end{cases} \qquad (12)$$

Uma vez encontrado, é possível determinar α_i w através da equação principal. Existem três propriedades importantes: Primeiro, a expressão é convexa. Em segundo lugar, os dados entram na expressão apenas na forma de produtos internos de pares de exemplos. Por fim, com determinado, nota-se que a maioria é zero, mas aqueles com valores diferentes de zero indicam que o correspondente é um vetor de suporte.$\alpha \ \alpha_i \ x_i$

A segunda propriedade também vale para a equação discriminante correspondente à Eq. (13).

Revista ARACÊ
ISSN: 2358-2472

$$h(x) = \text{sign}(\sum_l a_l y_l (x_l \cdot x) + b) \qquad (13)$$

Os vetores de suporte são os exemplos que satisfazem e estão localizados na borda. Este fato pode ser usado para calcular x_i $y_i(w \cdot x_i + b) = 1$ b a partir de qualquer vetor de suporte usando Para estabilidade numérica, isso deve ser feito para todos os vetores de suporte, tomando o valor médio de $b = y_i - w \cdot x_i$. b. A função discriminante resultante é chamada de Máquina de Vetores de Suporte.

Um dos maiores desafios, em máquinas de aprendizado estatístico, diz respeito a encontrar um kernel que otimize o limite de decisão entre as classes de uma determinada aplicação. Nas redes neurais ELM, um kernel linear, por exemplo, é capaz de resolver um problema linearmente separável, como visto em 5 (a). Seguindo o mesmo raciocínio, os kernels sigmoide, RBF e sinusoidal são capazes de resolver problemas separáveis pelas funções sigmoidal, radial e sinusoidal, vistas na Fig. 5 (b), na Fig. 5 (c) e na Fig. 5 (d), respectivamente.

Dessa forma, a capacidade de generalização de uma boa rede neural pode depender de uma escolha bem ajustada do kernel. O melhor kernel pode estar subordinado ao problema que está sendo resolvido. Como efeito colateral, investigar diferentes kernels geralmente é um processo caro que envolve validação cruzada combinada com diferentes condições iniciais aleatórias. No entanto, a investigação de grãos distintos pode ser necessária; caso contrário, a rede neural composta, por um kernel mal ajustado, pode gerar resultados insatisfatórios. Como contra-exemplo, observe o emprego do kernel linear aplicado às distribuições sigmóide e senoidal apresentadas na Fig. 6 (a) e na Fig. 6 (b), respectivamente. As precisões de classificação expostas na Fig. 6 (a) e na Fig. 6 (b) são de 78,71% e 73,00%, respectivamente. Visualmente, é possível observar que o kernel Linear não mapeia adequadamente os limites de decisão das distribuições Sigmoide e Sinusoidal.

Uma boa capacidade de generalização desses kernels também depende de uma escolha bem ajustada de parâmetros (C, γ). O parâmetro de custo C refere-se a um ponto de equilíbrio razoável entre a largura da margem do hiperplano e a minimização do erro de classificação referente ao conjunto de treinamento. O parâmetro kernel γ controla o limite de decisão em função das classes (Lima, 2021, 2022a, 2022b 2023). Não existe um método universal no sentido de escolher os parâmetros (C, γ). No presente trabalho, os parâmetros C e γ variam exponencialmente em sequências crescentes, matematicamente de acordo com a função , onde 2^nn = {0, 5, 10}. A hipótese é verificar se esses parâmetros distintos dos padrões; (C, γ) = (,) são capazes de gerar melhores precisões.$2^0 2^0$

Revista ARACÊ
ISSN: 2358-2472

Figura 5: Desempenhos bem-sucedidos de *kernels* compatíveis com os conjuntos de dados, esses são exemplos autorais hipotéticos de educação

Revista ARACÊ
ISSN: 2358-2472

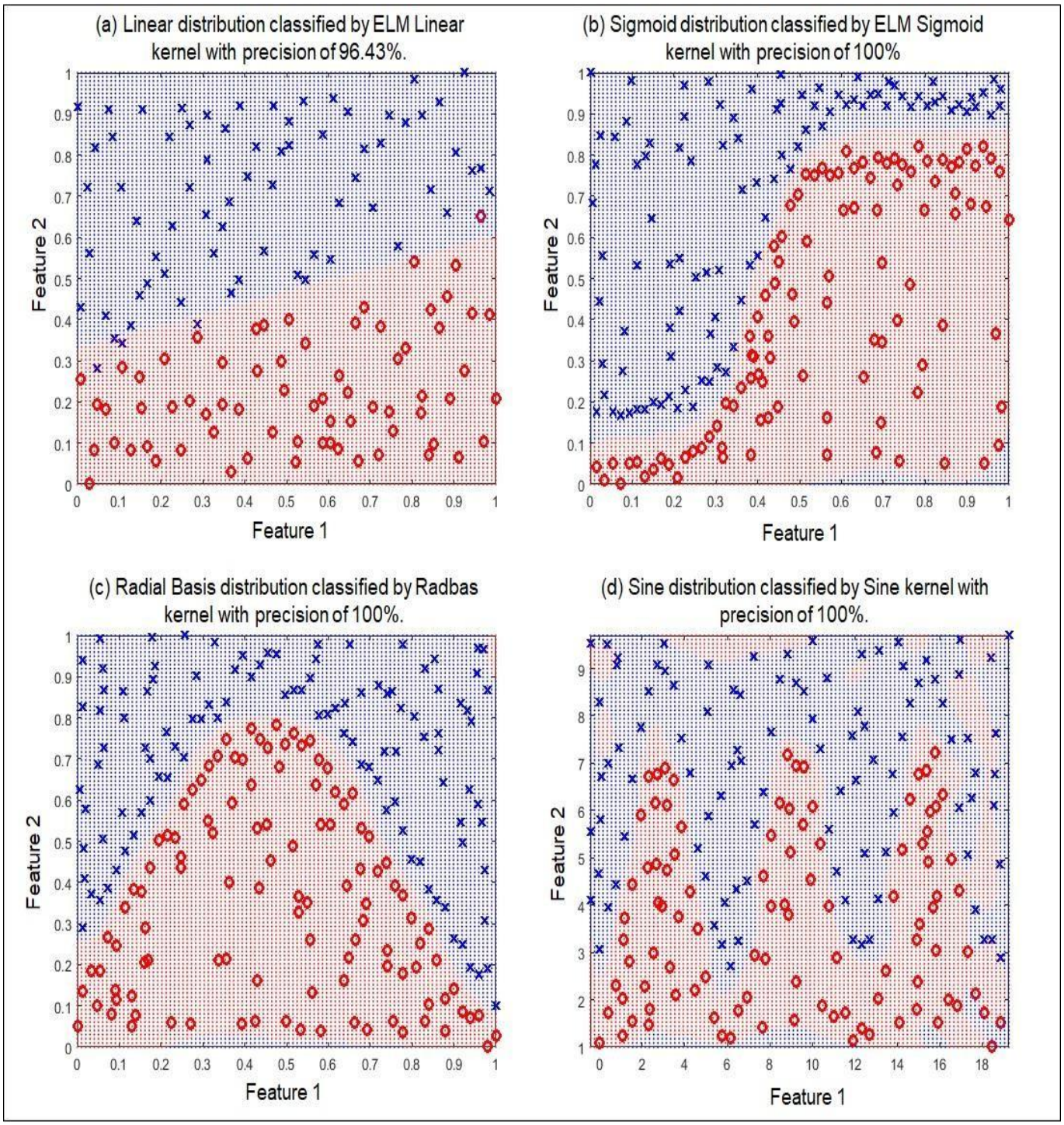
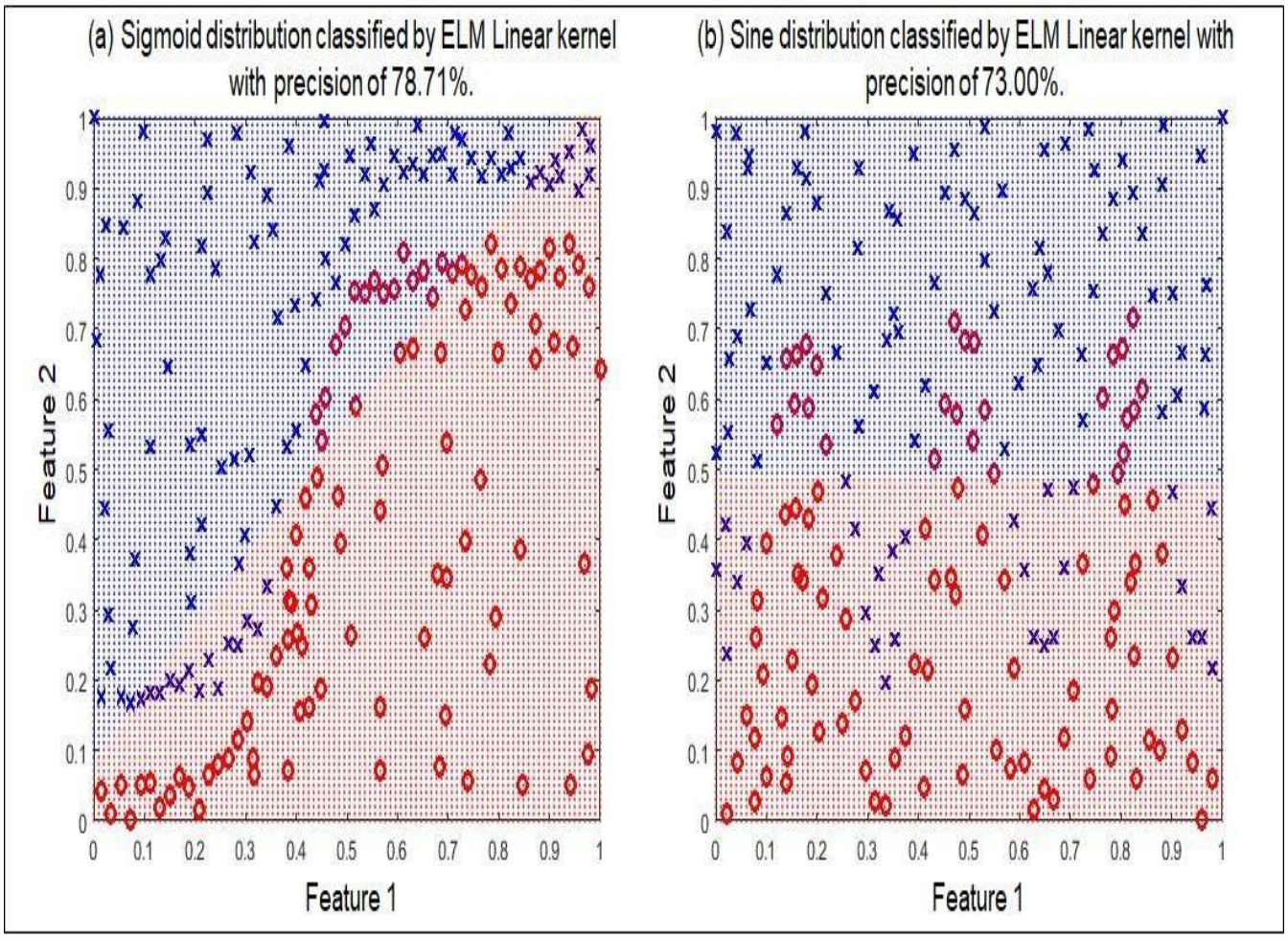


Figura 6: Desempenho malsucedido do *kernel linear* em conjuntos de dados não separáveis linearmente. Estes são exemplos dedutivos hipotéticos autorais

Revista ARACÊ
ISSN: 2358-2472



3 MATERIAIS E MÉTODOS

O presente trabalho utiliza dois cenários voltados para o reconhecimento de padrões de clusters pertencentes à classe alvo. No primeiro cenário, aqui denominado "simples", a classificação é binária. Há apenas classe (JPEG) vs. contraclasse (PNG). Essa metodologia foi desenvolvida por Pavel (2017) e replicada no cenário simples acima mencionado.

Em um segundo cenário, denominado "complexo", foi utilizado o método um-contra-todos. De acordo com a Tabela 3, a classe-alvo (JPEG) está dispersa em uma miscelânea de contraclasses. A criação do cenário complexo visa estabelecer uma simulação da máquina de um determinado usuário comum. Esse usuário tem uma tendência a ter uma grande quantidade de arquivos para as mais diversas finalidades. O objetivo do trabalho é aproximar o experimento do uso comum de computadores pessoais pela sociedade contemporânea.

Todos os arquivos estão disponíveis gratuitamente no repositório de direitos autorais (Dejavu, 2024). O objetivo é que o trabalho proposto possa ser replicado por terceiros. Visa também garantir a veracidade dos resultados alcançados pela metodologia proposta.

Revista ARACÊ
ISSN: 2358-2472

3.1 CENÁRIO SIMPLES: CLASSIFICAÇÃO BINÁRIA

No cenário apresentado, foram designados 1.000 (mil) arquivos JPEG e 1.000 (mil) arquivos PNG, considerados respectivamente classe e contraclasse. Os arquivos foram catalogados na rede mundial de computadores.

O trabalho proposto cria uma ferramenta autoral visando o reconhecimento de padrões de classe alvo para arquivos JPEG. A partir de um conjunto de treinamento, é possível formular uma hipótese sobre os clusters da classe alvo (JPEG). Cabe ao sistema autoral estimar a classe de um cluster inédito. Ele faz isso comparando suas características auditadas em tempo real e aquelas obtidas durante a etapa de treinamento, A intenção é que o trabalho proposto possa ser replicado por outros pesquisadores, fortalecendo a verificabilidade dos resultados alcançados pela metodologia proposta. Todos os experimentos foram realizados em um computador equipado com um processador Intel(R) Core(TM) i5-11400H de 11ª geração, com velocidade variando de 2,70 GHz a 2,69 GHz. Há 40.0 GB de RAM, 2 TB. Há uma placa de vídeo GeForce GTX 1650.

3.2 CENÁRIO COMPLEXO: CLASSIFICAÇÃO DE MÉTODO UM CONTRA TODOS

Nesse cenário, foi formado um banco de dados contendo 16 mil arquivos, mil de cada extensão apresentada a seguir. Os arquivos foram catalogados na rede mundial de computadores. O principal objetivo do experimento é simular o computador de um usuário comum. Na contemporaneidade, os computadores eletrônicos, como computadores de mesa, telefones celulares e outros dispositivos, não são focados apenas no entretenimento. Tornaram-se indispensáveis para o cumprimento de tarefas domésticas, profissionais, acadêmicas e de planejamento.

A hipótese é que o sistema autoral é capaz de diferenciar clusters da classe alvo (JPEG) de outros tipos de arquivos, mesmo que eles não tenham sido apresentados durante sua fase de treinamento. O método um-contra-todos é aplicado neste cenário complexo. Em termos estatísticos, só importa a segregação dos clusters da classe alvo em detrimento de todas as outras classes.

Tabela 1 - Cenário Complexo: Conjunto de Dados Autorais.

Classe	Contraclasse	Descrição
JPEG	PNG	Formatos de imagem comumente usados para imagens digitais.
Word (doc/docx)	Escritor - LibreOffice (ODT)	Editores de texto para criar documentos, papéis, etc.

Revista ARACÊ
ISSN: 2358-2472

Excel (xls/xlsx)	Calc - LibreOffice (ODS)	Programas de planilhas para gerenciamento de dados e cálculos.
PowerPoint (ppt/pptx)	Impress - LibreOffice (ODP)	Aplicativos para criação de apresentações para seminários, projetos, etc.
Acesso (accdb)	SQL	Sistemas de gerenciamento de banco de dados.
Outlook (msg)	Gmail (EML)	Formatos de arquivo de e-mail usados para comunicação.
OneNote (um)	Joplin (JEX)	Cadernos virtuais para organizar notas e tarefas.
Editora (pub)	TIFF	Ferramentas para criar folhetos, cartões e outras mídias visuais.

4 METODOLOGIA

4.1 EXPERIMENTOS VOLTADOS PARA A RECUPERAÇÃO DE DADOS FORMATADOS

Todos os arquivos foram alocados em um dispositivo de memória flash - pendrive, e uma esterilização lógica prévia foi realizada neste dispositivo. O comando fdisk lista as partições reconhecidas pelo sistema operacional.

Se a partição não estiver listada, a análise forense digital será inviável em nossos materiais e métodos. Isso se deve às ferramentas das distribuições Linux para análise forense digital. Por exemplo, a ferramenta nativa de coleta de dados - dd: ferramenta nativa de coleta de dados como Kali e Parrot Linux - em distribuições Linux só funciona se o sistema operacional listar a partição.

Os principais problemas relacionados à partição não listada correspondem ao driver não ser reconhecido. Outra possibilidade é que o dispositivo esteja parcial ou completamente quebrado. Em seguida, a esterilização lógica reorganiza o dispositivo para definir o sistema de arquivos, ou seja, funciona como limpar os diretórios do dispositivo. Este procedimento foi realizado usando o comando wipe:

- Para localizar o diretório do dispositivo:

fdisk –l

- Processo de esterilização:

wipe/dev/sda

Após a esterilização lógica, todas as amostras de classe versus contraclasse são copiadas para o dispositivo de memória. Posteriormente, é realizada uma formatação simples, que não considera a esterilização lógica. Este procedimento pode ser realizado através da própria exclusão por interface gráfica.

Na sequência, há a coleta de dados. Essa coleta corresponde à duplicação forense, ou seja, uma cópia bit a bit do conteúdo encontrado no dispositivo que está sendo investigado. O arquivo com a

Revista ARACÊ
ISSN: 2358-2472

extensão dd. que se refere à imagem do dispositivo de destino.

Antes do tribunal, a perícia deve ser realizada nesta extensão que funciona como uma cópia dos dados. O dispositivo eletrônico deve ser devolvido ao réu o mais rápido possível. O réu deve ter amplo direito de defesa. Para isso, é necessário devolver seu equipamento eletrônico após copiar a imagem do dispositivo de destino.

Além disso, o uso constante do dispositivo de armazenamento pode reduzir sua vida útil parcial ou completamente. Esse esgotamento é conhecido como blocos defeituosos. O estresse promovido por uma investigação forense é enfatizado pela promoção de uma varredura bit a bit do artefato eletrônico. De acordo com a redução da vida útil do dispositivo, as possíveis evidências materiais de uma investigação podem ser perdidas ou comprometidas.

Exceções podem ocorrer, dependendo das interpretações do magistrado e da equipe técnica. Em alguns casos, as autoridades devem confiscar ou destruir um dispositivo. Isso ocorre em parte porque devolvê-lo seria uma ofensa devido ao seu material armazenado.

Para a coleta de dados, o Simpósio Brasileiro de Segurança da Informação recomenda o software dd: disk dump.

O Simpósio Brasileiro de Segurança da Informação recomenda o uso do software dd (disk dump) para coleta de dados. Com a adoção do dd, nosso objetivo é que nossa ferramenta seja utilizada pelas forças policiais e aceita pelo sistema judiciário brasileiro. Infelizmente, o Brasil e muitos países da América do Sul, como a Argentina, fazem parte de uma rota de dados maliciosos, como botnets (Research, 2023). Muitas atividades maliciosas são armazenadas em servidores brasileiros. Pretendemos fazer com que a prática forense brasileira reconheça nossa ferramenta adotando o dd. Para isso, o apoio do Simpósio Brasileiro de Segurança da Informação é fundamental.

Os parâmetros dd correspondem a:

- -if: partição investigada;
- -of: imagem coletada (cópia bit a bit) da partição investigada.
- No terminal, usamos:

dd if=/dev/sda1 of=imagem.dd

Deve-se notar que o dd não realiza detalhado, ou seja, não imprime nenhuma informação no terminal. Para o usuário, aparentemente, resulta na sensação de que houve uma falha no processamento. Após a coleta, as seguintes ferramentas de recuperação de dados formatadas são empregadas.

Revista ARACÊ
ISSN: 2358-2472

4.1.1 Primeiro

O Foremost é um software de terminal cujo objetivo é recuperar dados formatados através de Data Carving, considerando os cabeçalhos, rodapés e estruturas de arquivos, podendo trabalhar com arquivos de imagem ou diretamente em uma determinada unidade. Usamos a versão 1.5.7 do Foremost, lançada em 12/07/2012.

- -t: especifica o tipo de arquivos a serem recuperados, por exemplo, JPEG, gif, png, bmp, ou você pode usar all para recuperar todos os tipos de arquivos.
- -i: abreviação de input, para a partição de origem.
- -o: abreviação de saída, para o caminho de destino dos arquivos a serem recuperados (em outra partição, como um dispositivo de memória auxiliar). Por padrão, foremost cria uma pasta chamada output, se o usuário não definir o destino.
- No terminal, usamos o Foremost:

primeiro -t all -i image.dd

4.1.2 Bisturi

Por padrão, todos os tipos de arquivos estão contidos no arquivo de configuração padrão (arquivos systems/etc/scalpel/scalpel.conf). Esse arquivo contém comentários que correspondem ao padrão de configuração. Para especificar quais são os tipos de arquivos a serem extraídos, é necessário descomentar as linhas referentes às extensões dos arquivos. Usamos o Scalpel versão 1.60, lançado em 27/06/2013.

- A partição de origem. Observe que não há diretiva -i, embora a documentação do Scalpel afirme que ela é necessária.
- -o: abreviação de saída, para o caminho de destino dos arquivos a serem recuperados. Sua pasta deve primeiro ser criada antes que o Scalpel seja invocado.
- -c: arquivo de configuração discutido anteriormente. No terminal, usamos bisturi:

bisturi image.dd -o /saída2/

-c /etc/bisturi/bisturi-bisturi.conf

4.1.3 Resgate Mágico

O Magic Rescue também emprega o Data Carving para recuperar dados foscos. Por padrão, todos os tipos de arquivos a serem recuperados estão contidos em: (files systems/usr/share/magicrescue/recipes) Usamos o Magic Rescue versão 1.1.10, lançado em 24/11/2018.

Revista ARACÊ
ISSN: 2358-2472

- -d: para o caminho de destino dos arquivos a serem recuperados. É necessário que a pasta seja criada antes que o Magic Rescue seja invocado.
- Não há sinalizador (diretiva) para a partição de origem, basta colocá-lo entre aspas no comando.
- -r: arquivos de configuração discutidos anteriormente. No terminal, use o Magic Rescue:

magicrescue -d output3 image.dd -r avi
-r canon-cr2 -r elf -r flac -r gpl -r gzip
-r jpeg-exif -r jpeg-jfif -r mbox
-r mbox-mozilla-caixa de entrada -r mbox-mozilla-enviado
-r mp3-id3v1 -r mp3-id3v2 -r msoffice
-r nikon-raw -r perl -r png -r ppm
-r sqlite -r zip

4.1.4 Fotorecreação

PhotoRec é um aplicativo de código aberto. Tem a função de recuperar dados que não podem ser abertos, podendo ser utilizado em dispositivos móveis como pendrive, CDs e HDs. Usamos o PhotoRec versão 7.2, lançado em 22/02/2024.

4.1.5 Recuva

O Recuva permite recuperar arquivos que foram excluídos no sistema Windows. Essa recuperação não se restringe ao disco rígido, mas também possibilita o resgate de arquivos salvos em dispositivos portáteis. Sua principal função é localizar arquivos que podem ser recuperados. No entanto, o programa também permite uma exclusão completa dos arquivos. Usamos o Recuva versão 1.53.2096, lançado em 13/06/2023.

4.1.5 Autópsia

A autópsia possibilita a recuperação de dados excluídos. A autópsia possibilita a recuperação de dados excluídos. Com essa ferramenta, você não pode acessar diretamente a unidade ou a imagem para executar a Escultura de Dados. Você deve seguir o processo de criação do caso, construção do arquivo de visualização, os arquivos de resultados, análise dos resultados, execução de verificações de integridade e extração dos dados. Usamos a versão 4.21.0 do Autopsy, lançada em 06/09/2023.

Revista ARACÊ
ISSN: 2358-2472

4.1.6 Gênio do disco

O DiskGenius é usado para recuperar arquivos excluídos, perdidos ou formatados de uma variedade de dispositivos de armazenamento, incluindo discos rígidos, HDDs externos, unidades flash USB, discos virtuais, cartões de memória e matrizes RAID. Usamos o DiskGenius versão 5.6.1.1580, lançado em 15/08/2024.

4.1.7 Deca

O DECA emprega aprendizado de máquina para reconhecimento de padrões de clusters em arquivos JPEG (Gladyshev; Tiago, 2017). Na fase de extração de recursos, o DECA constrói um histograma do cluster avaliado. Então, na fase de classificação, o histograma de cluster serve como neurônios de entrada para a máquina de aprendizado estatístico. A DECA não emprega reconhecimentos de padrões distintos para identificar cabeçalhos e rodapés. O aprendizado de máquina é único.

Em relação ao reconhecimento de padrões, o DECA assume um kernel linear. O kernel mencionado funciona de forma eficaz quando as distribuições são linearmente separáveis. Portanto, como método de análise, não se aprofunda nas diferentes funções de aprendizagem. Diferentes parâmetros de custo e variações de parâmetros do kernel não são explorados.

Parâmetros da ferramenta DECA:

- -vv: abreviação de verbose, que imprime o progresso do exame na tela.
- -o: abreviação de saída, para o caminho de destino dos arquivos a serem recuperados (em outra partição, como um dispositivo de memória auxiliar). É necessário que a pasta seja criada antes que o DECA seja invocado.
- Não há marcação (diretiva) para a partição de origem, apenas mencione-a no comando.
- –linear: sem aprendizado de máquina, apenas Data Carving.
- –deca: com aprendizado de máquina adicionado ao Data Carving.
- -m jpeg.model: arquivo contendo os parâmetros de configuração do aprendizado de máquina.
- No terminal, usamos o DECA da seguinte forma:

./deca -vv -o /home/kali/Desktop/saída

--deca -m jpeg.model image.dd

4.1.8 Dejavu Forense: Técnica Proposta pelos Autores

O Dejavu Forensics usa as seguintes bibliotecas:

Revista ARACÊ
ISSN: 2358-2472

- libtsk-dev (sleuthkit): responsável pela leitura de clusters do dispositivo de armazenamento de destino.
- libmagic-dev: Gerenciador de escultura de dados (números mágicos).
- liblinear-dev: responsável pela etapa de reconhecimento de padrões usando discriminante linear.
- libsvm-dev: responsável pelo estágio de reconhecimento de padrões usando a Support Vector Machine.

Tanto o Autopsy quanto o Dejavu Forensics usam o Sleuth Kit para processar os clusters de uma partição. Mas o grande diferencial do Dejavu é sua integração com o aprendizado de máquina. O Sleuth Kit é vital para a recuperação de dados em ambas as ferramentas. Mas Dejavu vai além. Ele usa aprendizado de máquina para recuperar dados e encontrar padrões nos clusters. Essa abordagem permite uma recuperação mais precisa e eficiente. Ele difere das técnicas tradicionais por usar aprendizado de máquina no processo de recuperação. Isso o torna compatível com a análise baseada em padrões.

A técnica sugerida usa aprendizado de máquina para detectar padrões de cluster em arquivos JPEG, especificamente usando o Support Vector Machine (SVM). Na fase de extração de recursos, a ferramenta cria um histograma do cluster analisado. Posteriormente, durante a etapa de classificação, esse histograma atua como um conjunto de recursos de entrada para o modelo de aprendizado estatístico.

A base dessa ferramenta é baseada na metodologia DECA, um algoritmo de Data Carving que combina a identificação de assinaturas digitais específicas, conhecidas como "números mágicos", e o reconhecimento de padrões de cluster por meio de aprendizado de máquina. O DECA foi originalmente projetado para recuperar com eficiência dados JPEG não fragmentados. Isso envolve a identificação de clusters com dados JPEG, feita detectando suas assinaturas digitais de cabeçalho e rodapé.

Além disso, a ferramenta Dejavu usa uma estratégia de aprendizado de máquina para reconhecer padrões de arquivo JPEG. Nesse contexto, é empregada a Máquina de Vetores de Suporte (SVM), que tem se mostrado eficaz no reconhecimento de padrões em conjuntos de dados complexos. A classificação SVM funciona com base no princípio de criar uma separação ideal entre as classes. O principal objetivo do SVM é estratificar os dados criando uma superfície de decisão ideal, chamada de hiperplano. Esta superfície de decisão visa otimizar a precisão da classificação do conjunto de treinamento, garantindo as melhores margens em relação aos vetores de suporte. A ideia subjacente é que o hiperplano ideal terá uma melhor capacidade de generalização quando aplicado ao conjunto de testes.

Revista ARACÊ
ISSN: 2358-2472

O Dejavu Forensics usa os seguintes parâmetros:

- -vv: Esta opção significa detalhado. Ele exibe o progresso.
- -o: Esta opção significa saída, mostrando onde salvar os arquivos recuperados. A pasta de destino deve ser criada antes de executar o comando.
- –dejavu: essa opção adiciona aprendizado de máquina ao processo de Escultura de Dados.
- -oneclass: Isso significa que a análise será feita em um único modo ou para uma classe de arquivo.
- -fex "raw": Isso diz que o método de extração de recursos usado é bruto, referindo-se à extração direta de dados de arquivos. histo pode ser escolhido quando os recursos de entrada são sobre o histograma do cluster.
- /dev/sdb1: Este é o dispositivo ou partição a ser verificado, neste caso, /img.dd.
- No terminal, usamos o Dejavu da seguinte forma:

./dejavu -vv -o /home/kali/Desktop/Dejavu/output
-oneclass -fex "bruto" img.dd

Esses comandos mostram diferentes configurações e abordagens na ferramenta Dejavu para recuperação de arquivos PNG e JPG. Além disso, a opção sem rodapé controla se o aprendizado de máquina é usado para reconhecer padrões de rodapé nos resultados da análise. Registramos a primeira patente para Dejavu em 26/07/2023.

O Dejavu difere da ferramenta DECA existente de várias maneiras. O Dejavu pode identificar arquivos PNG e JPG, enquanto o DECA só pode detectar arquivos JPG. No reconhecimento de padrões, o DECA usa apenas o kernel linear. Como mostra a Figura 6, o kernel linear tem um desempenho ruim em distribuições separáveis não linearmente. Por outro lado, o Dejavu explora uma variedade de kernels, incluindo linear, polinomial, sigmoidal e radial. Para cada kernel, investigamos diferentes condições iniciais.

Estes incluem parâmetros de kernel e gama (relacionados à curvatura). Nossa estrutura também explora diferentes métodos de extração de recursos. Isso inclui o cluster bruto e o histograma do cluster. O DECA só pode analisar o histograma do cluster. Embora o Dejavu Forensics tenha sido desenvolvido com base na metodologia proposta pela ferramenta DECA (2017), não nos limitamos a ela. Expandimos a abordagem para recuperar dados formatados usando aprendizado de máquina, o que nos permitiu obter resultados superiores em termos de precisão e tempo de recuperação. Optamos por comparar o Dejavu Forensics com ferramentas comerciais populares, como Recuva, Autopsy e DiskGenius. Isso mostra sua relevância hoje. Nosso objetivo era fazer mais do que apenas testar ferramentas acadêmicas. Além disso, a seção 6. O documento inclui uma comparação com ferramentas

Revista ARACÊ
ISSN: 2358-2472

comerciais de recuperação de dados. Ele detalha os resultados e destaca o desempenho superior da Dejavu Forensics.

5 RESULTADOS

Dada a importância da recuperação de arquivos de dispositivos formatados, este estudo tem como objetivo examinar o desempenho e as limitações do Data Carving, utilizado para recuperar dados de dispositivos formatados, o que envolve a identificação das assinaturas iniciais (cabeçalhos) e terminações (rodapés) associadas às extensões de arquivo correspondentes.

A premissa subjacente é que cada tipo de arquivo tem uma estrutura característica de bytes no início e no final. No entanto, a abordagem baseada na identificação de cabeçalhos e rodapés, chamada de "número mágico", pode levar a desafios, tais como: (i) a geração de um grande número de falsos positivos, ou seja, arquivos inexistentes; (ii) a exclusão acidental de arquivos existentes; e (iii) a recuperação de arquivos corrompidos contendo apenas fragmentos de dados.

O uso de aprendizado de máquina surge como uma alternativa promissora para superar as limitações do Data Carving. Embora o aprendizado de máquina seja amplamente conhecido e aplicado em diversos campos computacionais, seu uso na área forense digital ainda está em um estágio inicial.

As Tabelas 2, 3, 4 e 5 apresentam os resultados obtidos por diferentes instrumentos de análise. recuperação de dados formatados.

As Tabelas 4 e 5 mostram os resultados para recuperar arquivos no formato PNG. Observe que o DECA não foi incluído nesses resultados. O DECA só pode detectar arquivos JPEG, como descrevem as Tabelas 2 e 3.

As ferramentas utilizadas foram Foremost, Scalpel, Magic Rescue, Photorec, Recuva, DiskGenius, Autopsy, DECA e Dejavu Forensics, desenvolvidas no presente estudo. Embora as ferramentas de última geração sejam funcionais, elas não empregam aprendizado de máquina. Por outro lado, a Dejavu Forensics emprega o aprendizado de máquina como uma abordagem para superar as limitações da escultura de dados.

Ao analisar as tabelas, os termos "N mágicos" referem-se a "números mágicos". São sequências de bytes que indicam o formato dos arquivos. Enquanto o termo "ML" (Machine Learning ou Learning of Machine) é usado para identificar ferramentas que usam essa abordagem para recuperação de dados.

A Tabela 2 apresenta os resultados obtidos por cada software no Cenário Simples. Restrito a arquivos JPEG e PNG formatados. O Scalpel teve um problema preocupante, com uma taxa de arquivos falsos positivos de 99,90%, o que significa que, entre os arquivos recuperados, apenas 1 era realmente verdadeiro e os outros foram recuperados errados. O Magic Rescue também teve uma alta

Revista ARACÊ
ISSN: 2358-2472

porcentagem de arquivos falsos positivos, chegando a 96,08%, e ainda gerou arquivos repetidos.

Quanto à geração de arquivos repetidos durante o processo, a Autópsia se destacou negativamente nesse aspecto. Embora tenha alcançado com 100% de precisão em relação aos dados originais, o Autopsy gerou quase o dobro do número de arquivos repetidos, o que resultou em 999 arquivos a mais do que os dados originais. Portanto, o Autopsy gerou 49,97% de arquivos duplicados, embora verdadeiros e pertencentes ao banco de dados de dados. Ao estabelecer uma relação entre o tempo de execução e a recuperação do verdadeiro positivo, o Dejavu Forensics recupera 100% dos arquivos formatados em apenas 7 segundos.

DiskGenius, nos cenários mais simples, a ferramenta mostrou-se eficaz, principalmente para arquivos JPEG, onde alcançou uma precisão de 80,10%. No entanto, no cenário mais simples para arquivos PNG, a ferramenta teve um desempenho menos impressionante, com uma precisão de 40,70% em relação ao banco de dados e 76,79% em arquivos verdadeiros. Isso foi acompanhado por uma taxa de falsos positivos de 7,55% e uma taxa de duplicação de 15,66%.

As Tabelas 3 e 5 apresentam os resultados obtidos no Cenário Complexo. Existem 16.000 arquivos com 16 extensões amplamente utilizadas por usuários comuns. Quando olhamos para a possibilidade de resultados falsos positivos, vemos que o Dejavu e o Recuva não geraram nenhum arquivo falso positivo. Isoladamente, destaca-se a ferramenta Dejavu, que, além de garantir uma taxa de recuperação total de 100% dos arquivos, realizou a operação em um tempo notável de 13 segundos. A Dejavu Forensics não tinha conhecimento de nenhuma manifestação de arquivos falsos positivos e/ou repetidos.

A Tabela 4 refere-se à recuperação de dados no formato PNG no cenário simples. A análise das ferramentas revelou que a maioria das ferramentas atingiu níveis de precisão acima de 96%, excluindo o Scalpel, que não conseguiu recuperar nenhum arquivo. Em relação à geração de arquivos repetidos, apenas Magic Rescue e Autopsy geraram duplicatas, registrando proporções de 42,86% e 49,85%, respectivamente. Esse achado denota que, apesar da acurácia aceitável, os dois softwares geraram quase metade dos arquivos de dados de forma repetida.

Em relação ao tempo de processamento, mais uma vez, o Dejavu se destaca, sendo capaz de recuperar dados formatados em apenas 9 segundos. Vale destacar que, no cenário em análise, a Dejavu Forensics demonstrou uma precisão de 98,27%, com a incidência de 1,73% de arquivos gerados erroneamente, mostrando-se uma alternativa eficaz e ágil.

Na Tabela 5, são apresentados os resultados alcançados no cenário complexo, compreendendo um amálgama de 16 mil arquivos que visa emular o comportamento de um usuário comum. Em termos de geração de arquivos repetidos, o Foremost gerou 17,75% dos arquivos repetidos, enquanto o Magic

Revista ARACÊ
ISSN: 2358-2472

Rescue teve uma taxa de repetição de 20,55%. Mais uma vez, o Autopsy chama a atenção porque, apesar de ostentar uma precisão próxima a 99,90 na recuperação de arquivos verdadeiros, apresentou uma taxa de 60,65% de arquivos repetidos. Mesmo em um cenário que inclui 16 mil arquivos, o Dejavu conseguiu resgatar 98,10% dos arquivos de extensão PNG.

De modo geral, as ferramentas Recuva e Photorec chamam a atenção pela acurácia alta e baixa incidência de resultados falso-positivos e repetidos. Além disso, a ferramenta Dejavu Forensics se destaca por sua precisão na recuperação de arquivos e, principalmente, por sua capacidade de ser eficiente em um período de tempo extremamente curto em comparação com outras ferramentas. A Dejavu Forensics estabelece o emprego de aprendizado de máquina e dados científicos como um caminho promissor no campo da análise forense digital.

Tabela 2 - Cenário simples do JPEG: resultados da ferramenta.

Ferramenta	Comp. Tipo	Nº de arquivos gerais	Runtime	Tamanho da dir. (MB)	Posição falsa (%)	Duplicatas verdadeiras (%)	Posição verdadeira de db (%)	Posição verdadeira (%)
Dejavu Forense	Números mágicos + aprendizado de máquina	1000	7 s	39.9 MB	0	0%	100%	100%
Deca	Números mágicos + aprendizado de máquina	1005	18,96 s	40.1 MB	0.10%	1.39%	99%	98.51%
Recuva	Números mágicos	996	5m 12s	39.8 MB	9%	0%	90.60%	91%
Primeiro	Números mágicos	998	10m 53s	39.8 MB	0,20%	0%	99.60%	99.80%
Bisturi	Números mágicos	1000	8m 17s	4600 MB	99.90%	0%	0.10%	0.10%
Resgate Mágico	Números mágicos	1555	7m 22s	202.5 MB	96,08%	1.16%	4.30%	2.77%
Fotorecreação	Números mágicos	999	4m 39s	39.9 MB	0%	0%	99.90%	100%
Autópsia	Números mágicos	1999	34m 17s	584.6 MB	0%	49.97%	100%	50.03%
DiskGenius	Números mágicos	831	2m 45s	258 MB	3,13%	0,48%	80.10%	96,39%

Tabela 3 - Cenário complexo JPEG: resultados da ferramenta

Ferramenta	Comp. Tipo	Nº de arquivos gerais	Runtime	Tamanho da dir. (MB)	Posição falsa (%)	Duplicatas verdadeiras (%)	Posição verdadeira de db (%)	Posição verdadeira (%)
Dejavu Forense	Números mágicos + aprendizado de máquina	1000	13 s	39.9 MB	0%	0%	100%	100%
Deca	Números mágicos +	1002	33,21 s	40 MB	0.20%	0%	100%	99.80%

Revista ARACÊ
ISSN: 2358-2472

	aprendizado de máquina							
Recuva	Números mágicos	1000	1h 14 m 21s	39.9 MB	0%	0%	100%	100%
Primeiro	Números mágicos	4902	11m 39s	403.1 MB	61.87%	17.75%	99.90%	20.38%
Bisturi	Números mágicos	412	13m 57s	1400 MB	100%	0%	0%	0%
Resgate Mágico	Números mágicos	798	2h 43m	51.4 MB	78.32%	20.55%	0.90%	1.13%
Fotorecreação	Números mágicos	1000	6 m 34s	40.1 MB	1.80%	0%	98.20%	98.20%
Autópsia	Números mágicos	2695	2h 27min 12s	108.5 MB	0.04%	62.86%	100%	37.11%
DiskGenius	Números mágicos	830	3min 28s	24.8 MB	0,12%	0%	82.90%	99.88%

Tabela 4 - PNG Cenário simples: resultados da ferramenta.

Ferramenta	Comp. Tipo	Nº de arquivos gerais	Runtime	Tamanho da dir. (MB)	Posição falsa (%)	Duplicatas verdadeiras (%)	Posição verdadeira de db (%)	Posição verdadeira (%)
Dejavu Forense	Números mágicos + aprendizado de máquina	985	9 s	252.4 MB	1.73%	0%	96.8%	98.27%
Recuva	Números mágicos	1000	5m 12s	252.4 MB	0.10%	0%	99.90%	99.90%
Primeiro	Números mágicos	973	10m 53s	210.5 MB	0.31%	0%	97.00%	99.69%
Bisturi	Números mágicos	0	8m 17s	0	0%	0%	0%	0%
Resgate Mágico	Números mágicos	1829	7m 22s	456.6 MB	2.79%	42.86%	99.40%	54.35%
Fotorecreação	Números mágicos	998	4m 39s	252.4 MB	0.30%	0%	99.50%	99.70%
Autópsia	Números mágicos	1998	34m 17s	504.8 MB	0.15%	49.85%	99.90%	50.00%
DiskGenius	Números mágicos	530	1m 9s	8.68 MB	7.55%	15.66%	40.70%	76.79%

Tabela 5 - PNG Cenário complexo: resultados da ferramenta.

Ferramenta	Comp. Tipo	Nº de arquivos gerais	Runtime	Tamanho da dir. (MB)	Posição falsa (%)	Duplicatas verdadeiras (%)	Posição verdadeira de db (%)	Posição verdadeira (%)
Dejavu Forense	Números mágicos + aprendizado de máquina	1010	13 s	260.4 MB	2.77%	0.10%	98.10%	97.13%
Recuva	Números mágicos	1001	1h 14m 21s	252.5 MB	0.20%	0%	99.90%	99.80%
Primeiro	Números mágicos	4410	11m 39s	365.2 MB	38.96%	39.84%	93.50%	21.20%
Bisturi	Números mágicos	539	13 m 57s	4500 MB	82.75%	17.25%	0%	0%

Revista ARACÊ
ISSN: 2358-2472

Resgate Mágico	Números mágicos	8060	2h 43m	83.3 MB	8.10%	90.58%	10.60%	1.32%
Fotorecreação	Números mágicos	1000	6m 34s	252.5 MB	0.50%	0%	99.50%	99.50%
Autópsia	Números mágicos	2554	2h 27m 12s	632.1 MB	0.26%	60.65%	99.90%	39.12%
DiskGenius	Números mágicos	527	1m 58s	8.47 MB	20.87%	3.04%	40.10%	76.09%

5.1 COMPARAÇÃO COM FERRAMENTAS COMERCIAIS DE RECUPERAÇÃO DE DADOS

A recuperação de dados formatados é uma das principais demandas no campo da perícia digital. Enquanto ferramentas populares como Recuva, Autopsy e DiskGenius oferecem soluções pagas, com empresas de capital aberto como a Gen Digital Inc. (proprietária da Recuva), que tem mais de 3.400 funcionários. O Dejavu Forensics surgiu como uma alternativa gratuita e de código aberto. Mesmo ao competir com empresas grandes e robustas como a Gen Digital, a Dejavu Forensics oferece desempenho superior por meio do uso de aprendizado de máquina (Dejavu, 2024).

Os principais desenvolvedores comerciais dessas ferramentas estão aprimorando-as. Suas atualizações podem ser semestrais, anuais ou até mais frequentes. A Seção 4 discute esse ciclo de atualização em detalhes. Abrange seus impactos e usos na perícia digital. Ele destaca como essas melhorias afetam o desempenho das ferramentas.

Ao contrário dos métodos tradicionais baseados apenas em File Carving, o Dejavu Forensics usa máquinas de vetores de suporte (SVM) para encontrar padrões de blocos e clusters de dados. A escultura de dados geralmente causa arquivos corrompidos ou recuperações falsas. Isso reduz significativamente a taxa de falsos positivos. Nos testes, a ferramenta recuperou mais de 96% dos arquivos PNG e JPEG. Ele superou as ferramentas pagas em precisão e velocidade, completando recuperações em menos de 13 segundos. Outro ponto crucial é a transparência e acessibilidade da ferramenta. O código-fonte do Dejavu Forensics está disponível para a comunidade, permitindo que seja replicado e adaptado por outros pesquisadores e profissionais da área.

6 PARÂMETROS DE CONFIGURAÇÃO SVM PARA OTIMIZAÇÃO NA RECUPERAÇÃO DE IMAGEM

Na recuperação de imagens, as pessoas usam Máquinas de Vetores de Suporte (SVM). Eles provaram ser eficazes (Gladyshev; Tiago, 2017). O SVM é um modelo de aprendizado de máquina supervisionado que pode ser aplicado para tarefas de classificação e regressão. Neste trabalho, o foco está na classificação de imagens. Tem como objetivo distinguir entre categorias predefinidas de conteúdo visual.

Revista ARACÊ
ISSN: 2358-2472

Nosso estudo explorou os parâmetros padrão da SVM. Também explorou uma ampla gama de parâmetros para otimizar a solução proposta. O objetivo era aumentar sua precisão. Realizamos uma investigação detalhada. Ele se concentrou na melhor maneira de definir os parâmetros SVM. A investigação analisou especificamente como recuperar arquivos formatados.

Em termos de extração de características, dois métodos foram explorados: o histograma do cluster e o próprio cluster bruto. No reino dos kernels, investigamos quatro funções distintas. Foram elas: Linear, Polinomial, RBF (Função de Base Radial) e Sigmoide. Os parâmetros (C, γ) têm variações exponenciais.

Eles aumentam de acordo com a função 2n, onde n = {0, 5, 10}. A hipótese é verificar se esses parâmetros, diferentes dos padrões; (C, γ) = , pode gerar melhores precisões. O repositório estatístico de aprendizagem estudou a classe versus anticlasse. Também estudou métodos de classe única. Ressalta-se que investigar o parâmetro de custo 2^0 , 2^0 C não faz sentido quando apenas uma única classe está contida no repositório. O parâmetro de custo C refere-se a um equilíbrio razoável entre a largura da margem do hiperplano durante a ponderação das classes. Quando há apenas uma classe, não há ponderação entre as diferentes classes. Para cada cenário, investigamos 400 configurações. Foram 400:2 extrações ∗ 4 kernels ∗ parâmetros de 5 C ∗ parâmetros de 5γ ∗ 2 tipos de repositórios.

A ferramenta Dejavu visa recuperar imagens JPG. Tanto para cabeçalho quanto para rodapé, o melhor parâmetro está na extração de recursos do cluster bruto. Está no espaço de busca de configurações otimizadas para recuperação de imagens JPG, tanto para cabeçalho quanto para rodapé. No modelo especialista em cabeçalho para arquivos JPG, o kernel RBF é usado. É conhecido por sua capacidade de lidar com recursos não lineares. Tem um parâmetro γ fixo de $2^5 = 32$.

Esse valor γ influencia diretamente a configuração do kernel. Ele determina a curvatura do radial RBF em relação aos dados. Para o modelo especialista em rodapé de arquivos JPG, o kernel otimizado foi Polynomial, grau 3, com um parâmetro γ fixado em . A melhor maneira de criar o repositório de aprendizado estatístico é para uma classe. Essa configuração geralmente é útil em cenários com um grande conjunto de dados de uma classe. O objetivo é detectar desvios ou anomalias do padrão de classe.$2^5 = 32$

A ferramenta Dejavu também visa recuperar imagens PNG. Ele manteve o melhor método de extração de recursos: o cluster bruto. No entanto, o valor γ foi alterado para . Este foi um grande aumento. Ele reflete a maior complexidade e variação das imagens PNG em comparação com o JPG. No modelo especialista em cabeçalho para arquivos PNG, foi utilizado o kernel RBF (Radial Basis Function). Quanto ao modelo especialista em rodapé de arquivos PNG, o kernel escolhido foi um polinômio grau 3. O kernel polinomial pode capturar padrões polinomiais nos dados. Este kernel é útil

Revista ARACÊ
ISSN: 2358-2472

nos casos em que as características das imagens e suas categorias têm uma relação não linear. Mas ainda pode ser modelado com polinômios. Os modelos de cabeçalho e rodapé têm a mesma metodologia na configuração do repositório. Considera uma única classe.$2^{10} = 1024$

Os parâmetros foram ajustados após uma série de experimentos e análises. Buscamos a melhor configuração para encontrar imagens de forma eficaz. A ferramenta Dejavu, usada para implementar o SVM, forneceu um meio robusto para testar e aplicar essas configurações. Os resultados alcançados mostram que as configurações dos parâmetros foram decisivas. Eles incrementaram o quão bem a recuperação de imagens funcionou para os formatos JPG e PNG.

7 CONCLUSÕES

Em uma sociedade pesada em tecnologia e saturada de mídia, este trabalho visa esclarecer crimes digitais. Ele se concentra na recuperação de dados formatados intencionalmente. Essa demanda surge à medida que enfrentamos crimes cibernéticos cada vez mais sofisticados. Ao formatar intencionalmente um disco ou dispositivo, a remoção de arquivos ocorre principalmente em uma esfera lógica, disponibilizando o espaço anteriormente ocupado para novos arquivos. No entanto, é crucial destacar que os dados originais ainda permanecem fisicamente localizados no dispositivo. Eles são estruturados em clusters, cada um com cabeçalhos diferentes que sinalizam o início de determinados arquivos ou extensões. Esses cabeçalhos e rodapés são de extrema importância para localizar os dados desejados no processo de recuperação.

A perícia forense, diante desse cenário, pode se beneficiar significativamente de mecanismos automatizados, especialmente com a incorporação de soluções inovadoras de Machine Learning. A ferramenta "Dejavu Forensics" é uma prova dessa inovação, demonstrando eficácia ao recuperar mais de 96% dos arquivos foscos, como PNG e JPEG, em questão de segundos.

O rigor científico-metodológico aqui utilizado sugere que a Dejavu Forensics pode oferecer contribuições substanciais para o avanço da perícia digital, proporcionando maior eficiência, precisão e confiabilidade nas investigações.

Todos os arquivos e dados utilizados neste trabalho estão no repositório (Dejavu, 2024). Isso apóia nosso compromisso com a transparência e a pesquisa replicável. Além disso, este estudo é relevante para mais do que apenas a recuperação de dados. Ele enfatiza a necessidade de tecnologia para combater os crimes digitais de hoje, como lavagem de dinheiro, evasão fiscal e pedofilia.

É possível afirmar que os objetivos desta pesquisa foram alcançados. Este trabalho não apenas apresenta uma inovação tecnológica significativa, mas também destaca a capacidade da tecnologia de servir aos direitos humanos, promovendo a justiça e melhorando a qualidade de vida na era digital. Em

Revista ARACÊ
ISSN: 2358-2472

última análise, serve como um farol, iluminando o caminho para futuras investigações e soluções na interseção de tecnologia, justiça e sociedade.

TRABALHO FUTURO

Nosso aprendizado de máquina foi treinado especificamente para os formatos png e jpeg. Esses formatos foram escolhidos devido à sua prevalência em investigações forenses digitais e seu uso generalizado em dispositivos eletrônicos. Focar o estudo nesses tipos de arquivo permitiu um melhor teste do método de aprendizado de máquina proposto. Eles são representativos de cenários típicos de recuperação de dados. Além disso, os arquivos JPEG e PNG são populares na ciência forense. Sua estrutura nos permite testar técnicas de aprendizado de máquina com eles.

Estamos cientes de que outros formatos de arquivo são importantes em investigações forenses, por isso planejamos expandir a ferramenta para outros tipos de arquivo em versões futuras. Nesta fase, nossa principal prioridade tem sido estabelecer a confiabilidade e certeza do modelo com os formatos mais comuns.

Embora nosso estudo tenha mostrado resultados promissores na recuperação de dados formatados nos formatos jpeg e png, reconhecemos que existem algumas limitações importantes. O método proposto só funciona em arquivos jpeg e png. Isso limita o uso da ferramenta Dejavu Forensics em outros tipos de arquivo com estruturas diferentes.

Devido aos bons resultados na recuperação de dados png e jpeg, expandiremos a ferramenta Dejavu Forensics. Ele cobrirá outros tipos de arquivos usados em perícia. Pretendemos treinar o modelo com um banco de dados contendo 16 tipos de arquivos, incluindo, além de jpeg e png, formatos como doc, odt, xls, ods, ppt, odp, accdb, sql, msg, gmail, one, jex, pub e tiff. Essa expansão permitirá que a ferramenta ofereça suporte a uma ampla gama de cenários de recuperação de dados, fornecendo uma solução mais rápida e robusta para casos envolvendo diferentes formatos de arquivo Treinar a ferramenta com esses 16 tipos de arquivo pode melhorar muito as investigações forenses. Isso ajudaria a polícia e outras agências a recuperar e analisar dados com mais rapidez e precisão. Uma pesquisa de arquivos melhor pode acelerar as investigações de crimes cibernéticos, como fraudes e violações de dados. Pode ajudar a encontrar evidências cruciais rapidamente.

Além disso, ao adicionar esses novos tipos de arquivo e expandir a ferramenta, esperamos um melhor uso do Dejavu Forensics em investigações complexas. A ferramenta deve ajudar em muitos casos. Isso inclui a recuperação de documentos comerciais, arquivos de imagem, mensagens, apresentações e bancos de dados. Isso ajudará na resolução de crimes digitais.

Outro objetivo futuro é estender a abordagem para a recuperação de dados formatados, mesmo

Revista ARACÊ
ISSN: 2358-2472

que estejam compactados ou incompletos. Esse avanço é fundamental para a segurança cibernética e a análise forense digital. Nesses casos, a preservação das evidências é vital.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

ALHERBAWI, N.; SHUKUR, Z.; SULAIMAN, R. Systematic literature review on data carving in digital forensic. Procedia Technology, v. 11, p. 86-92, mar. 2013.

ALI, R. R.; MOHAMAD, K. M.; JAMEL, S.; KHALID, S. K. A. A review of digital forensics methods for JPEG file carving. Journal of Theoretical and Applied Information Technology, v. 96, n. 18, p. 5841-5856, out. 2018.

CARRIER, B. File system forensic analysis. Boston: Addison-Wesley Professional, 2005.

CASEY, E. Digital evidence and computer crime: conducting digital investigations. 3. ed. Maryland: Elsevier, 2011.

DEJAVU. Dejavu forensics. Disponível em: https://github.com/DejavuForensics/version-1.0. Acesso em: mar. 2024.

DUDA, R. O.; HART, P. E.; et al. Pattern classification. New York: John Wiley & Sons, 2006.

GARFINKEL, S. Anti-forensics: techniques, detection and countermeasures. In: INTERNATIONAL CONFERENCE ON I-WARFARE AND SECURITY, 2., 2007, [S.l.]. Proceedings [...]. [S.l.]: [s.n.], 2007. p. 77-84.

GARFINKEL, S. L.; SHELAT, A. Remembrance of data passed: a study of disk sanitization practices. IEEE Security & Privacy, v. 1, n. 1, p. 17-27, jan. 2003.

GLADYSHEV, P.; JAMES, J. I. Decision-theoretic file carving. Digital Investigation, v. 22, p. 46-61, jun. 2017.

HANNAN, M. To revisit: what is forensic computing? In: AUSTRALIAN COMPUTER, NETWORK & INFORMATION FORENSICS CONFERENCE, 2004, [S.l.]. Proceedings [...]. [S.l.]: [s.n.], 2004. p. 103-111.

HAYKIN, S. Neural networks and learning machines. 3. ed. Upper Saddle River: Pearson, 2008.

INC., G. D. Perfil e dados da empresa - Gen Digital Inc. Disponível em: https://stockanalysis.com/quote/pra/GEN/company/. Acesso em: mar. 2024.

JAMES, D. Forensically unrecoverable hard drive data destruction. [S.l.]: Infosec Writers, 2006.

KENT, K.; CHEVALIER, S.; GRANCE, T. Guide to integrating forensic techniques into incident response. [S.l.]: Tech. Rep. 800-86, 2006.

LAURENSON, T. Performance analysis of file carving tools. In: IFIP TC 11 INTERNATIONAL CONFERENCE, SEC 2013, 28., 2013, Auckland. Security and privacy protection in information processing systems: proceedings [...]. Berlin: Springer, 2013. p. 419-433.

LIMA, S.; SILVA, S. H. M. T.; PINHEIRO, R. Next generation antivirus for javascript malware detection based on dynamic features. Knowledge and Information Systems, [S.l.], [s.n.], 2023.

Revista ARACÊ
ISSN: 2358-2472

LIMA, S.; SILVA, S.; PINHEIRO, R.; et al. Next-generation antivirus endowed with web-server sandbox applied to audit fileless attack. Soft Computing, [S.l.], [s.n.], 2022. Disponível em: https://doi.org/10.1007/s00500-022-07447-4. Acesso em: out. 2024.

LIMA, S.; SOUZA, D.; PINHEIRO, R.; SILVA, S.; et al. Next generation antivirus endowed with bitwise morphological extreme learning machines. Microprocessors and Microsystems, v. 81, 103724, dez. 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0141933120308693. Acesso em: out. 2024.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. Cambridge: Cambridge University Press, 2008.

HOSSEINZADEH, M.; RAHMANI, A. M.; B. V. M. B. M. M. M. Z. Improving security using SVM-based anomaly detection: issues and challenges. Springer-Verlag GmbH Germany, part of Springer Nature, [S.l.], [s.n.], 2020.

(OCC), D. C. O. Sextortion. Disponível em: https://occ.org.br/tipo_de_fraude/sextortion/. Acesso em: out. 2020.

PAL, A.; MEMON, N. The evolution of file carving. IEEE Signal Processing Magazine, v. 26, n. 2, p. 59-71, mar. 2009.

PINHEIRO, R.; LIMA, S.; SOUZA, D.; et al. Antivirus applied to JAR malware detection based on runtime behaviors. Scientific Reports - Nature, v. 12, 1945, 2022. Disponível em: https://doi.org/10.1038/s41598-022-05921-5. Acesso em: out. 2024.

RESEARCH, E. ESET takes part in global operation to disrupt the Grandoreiro banking trojan. Disponível em: https://www.welivesecurity.com/en/eset-research/eset-takes-part-global-operation-disrupt-grandoreiro-banking-trojan/. Acesso em: 17 set. 2024.

SARI, S. A.; MOHAMAD, K. M. A review of graph theoretic and weightage techniques in file carving. Journal of Physics: Conference Series, [S.l.], IOP Publishing, p. 052011, 2020.

SENCAR, H. T.; MEMON, N. Identification and recovery of JPEG files with missing fragments. Digital Investigation, v. 6, p. S88-S98, jan. 2009.

RUSSELL, S.; NORVIG, P. Artificial intelligence. [S.l.]: [s.n.], 2013.

VACCA, J. R. Computer forensics: computer crime scene investigation. [S.l.]: [s.n.], 2006.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J.; DATA, M. Practical machine learning tools and techniques. In: Data mining. [S.l.]: [s.n.], 2005.