v7.n12.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

LIMITES DA ANÁLISE DE METADADOS NA DETECÇÃO DE INDÍCIOS DE ADULTERAÇÃO EM ARQUIVOS PDF: RELATO DE CASO PERICIAL

LIMITS OF METADATA ANALYSIS IN DETECTING TAMPERING INDICATORS IN PDF FILES: A FORENSIC CASE REPORT

LÍMITES DEL ANÁLISIS DE METADATOS EN LA DETECCIÓN DE INDICIOS DE ADULTERACIÓN EN ARCHIVOS PDF: RELATO DE CASO PERICIAL

José Melquíades Neto[1], Mauro José Araújo de Melo[2], Raimundo Pereira da Cunha Neto[3]

https://doi.org/10.47820/recima21.v7i12.9756

PUBLICADO: 12/2026

RESUMO

Uma forma direta de automatizar o exame de comprovantes em PDF é procurar em cada arquivo sinais de edição posterior, como atualizações incrementais, datas divergentes e software de edição nos metadados. Essa abordagem pressupõe que o documento questionado é um documento verdadeiro alterado e falha quando ele é produzido do zero em processador de texto e exportado uma vez. Este relato de caso examina até onde os metadados sustentam essa detecção, com 21 documentos nativos digitais de um processo criminal e um parecer técnico como referência. Compararam-se duas versões de uma ferramenta própria: uma analisa cada arquivo isoladamente; outra confronta o software produtor com a origem alegada e com os demais arquivos do lote. A análise isolada deu aos legítimos escore médio maior que aos não conformes, por tratar como software de edição a biblioteca iText, padrão na emissão institucional; corrigido o erro, não alcançou os dois não conformes de estrutura mais limpa. O mesmo metadado, um processador de texto como criador, só se tornou indício quando confrontado com uma referência externa ao arquivo: a origem institucional alegada separou o lote, e a ferramenta predominante no lote pontuou seis dos sete não conformes e nenhum legítimo nos limiares de 1 a 15, mas nenhum no limiar de 20. O caso único, os critérios definidos após conhecido o resultado e um parecer também baseado em metadados, sem confirmação junto aos emissores, tornam o estudo exploratório. Os metadados de um arquivo, lidos isoladamente, indicam procedência, mas não bastam para afirmar autenticidade.

PALAVRAS-CHAVE: Metadados. Forense digital. Fraude documental. Perícia documental. Documentoscopia.

ABSTRACT

A direct way to automate the examination of PDF receipts is to look in each file for signs of later editing, such as incremental updates, diverging dates and editing software in the metadata. It assumes that a questioned document is an altered genuine one and fails when the document is produced from scratch in a word processor and exported once. This case report examines how far metadata support that detection, with 21 born-digital documents from a criminal case and an expert opinion as reference. Two versions of an in-house tool were compared: one analyses each file in isolation, the other checks the producing software against the claimed origin and the other files in the batch. The *isolated analysis gave legitimate documents a higher mean score than non-conforming ones, because it treated as editing software the iText library, standard in institutional issuance; once corrected, it still missed the two cleanest non-conforming documents.*

[1] Graduando em Engenharia de Software pelo iCEV – Instituto de Ensino Superior; atua em documentoscopia e forense digital.

[2] Mestre em Engenharia Elétrica; professor dos cursos de Engenharia de Software e ADM Tech do iCEV – Instituto de Ensino Superior.

[3] Mestre em Segurança de Redes pela UFMA e doutorando em Engenharia Biomédica pela UNIVAP; graduado em Ciência da Computação pela UESPI e especialista em Redes de Computadores; perito computacional forense.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

The same metadata item, a word processor as creator, became an indicator only when checked against a reference external to the file: the claimed institutional origin separated the batch, and the prevailing tool scored six of the seven non-conforming documents and no legitimate one at thresholds 1 to 15, but none at threshold 20. A single case, criteria set after the outcome was known and a reference opinion also based on metadata, unconfirmed with the issuers, make the study exploratory. The metadata of a file, read in isolation, indicate provenance, but are not sufficient to establish authenticity.

KEYWORDS: *Metadata. Digital forensics. Document fraud. Forensic document examination. Questioned documents.*

RESUMEN

Una forma directa de automatizar el examen de comprobantes en PDF es buscar en cada archivo señales de edición posterior, como actualizaciones incrementales, fechas divergentes y software de edición en los metadatos. Ese enfoque presupone que el documento cuestionado es un documento verdadero alterado y falla cuando se produce desde cero en un procesador de textos y se exporta una vez. Este relato examina hasta dónde los metadatos sostienen esa detección, con 21 documentos nativos digitales de un proceso penal y un dictamen técnico como referencia. Se compararon dos versiones de una herramienta: una analiza cada archivo aisladamente; otra confronta el software productor con el origen alegado y con los demás archivos. El análisis aislado dio a los legítimos puntuación media mayor que a los no conformes, por tratar como software de edición la biblioteca iText, estándar en la emisión institucional; corregido el error, no alcanzó los dos no conformes de estructura más limpia. El mismo metadato, un procesador de textos como creador, solo se volvió indicio al confrontarse con una referencia externa al archivo: el origen institucional alegado separó el lote, y la herramienta predominante puntuó seis de los siete no conformes y ningún legítimo en los umbrales de 1 a 15, pero ninguno en el umbral de 20. Caso único, criterios definidos tras conocer el resultado y un dictamen también basado en metadatos, sin confirmación de los emisores, hacen el estudio exploratorio. Los metadatos de un archivo, leídos aisladamente, indican procedencia, pero no bastan para afirmar autenticidad.

PALABRAS CLAVE: *Metadatos. Informática forense. Fraude documental. Pericia documental. Documentoscopia.*

1. INTRODUÇÃO

Comprovantes de pagamento, guias de recolhimento, boletos e extratos em formato PDF chegam aos autos em conjuntos de dezenas de arquivos, inclusive no processo penal. A lei penal tipifica a falsificação de documento público e a de documento particular (Brasil, 1940, arts. 297 e 298), e ambos os dispositivos preveem as mesmas duas condutas: falsificar o documento, no todo ou em parte, e alterar documento verdadeiro. Pune também o uso do documento falso (art. 304). No processo, a falsidade de documento constante dos autos pode ser arguida em incidente próprio ou verificada de ofício pelo juiz (Brasil, 1941, arts. 145 e 147). Por se tratar de infração que deixa vestígios, exige-se exame de corpo de delito (art. 158), a cargo de perito oficial (art. 159), facultado às partes apresentar parecer de assistente técnico (art. 159, § 5º, II). É nesse

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

ponto que o arquivo nativo digital chega ao perito, como no caso aqui relatado, que provém de processo criminal.

Uma forma direta de automatizar o exame desses arquivos é procurar, em cada PDF, vestígios de que ele foi alterado depois de criado: atualizações incrementais, datas de criação e de modificação divergentes e softwares de edição registrados nos metadados. A própria especificação do formato favorece essa leitura, ao prever a atualização incremental como o modo de modificar um arquivo existente sem reescrevê-lo (International Organization for Standardization, 2008, seção 7.5.6). Essa abordagem pressupõe que o documento questionado é um documento legítimo que foi editado. Há, porém, uma forma de falsificação que essa premissa não alcança: a fabricação integral, em que o documento é produzido do zero em processador de texto e exportado uma única vez (seção 2.2). A primeira versão da ferramenta aqui avaliada seguiu a abordagem de busca por vestígios de edição. Não se verificou em que medida as ferramentas de exame em uso a adotam; a primeira versão é tomada como representante dessa forma de exame, e a generalização de seus resultados é tratada como hipótese.

Soma-se a isso uma particularidade dos sistemas emissores. Parte dos sistemas bancários e governamentais gera documentos de forma automatizada no servidor, com a biblioteca de relatórios JasperReports acoplada à biblioteca iText (seção 2.4); é o caso dos dois emissores dos documentos legítimos aqui examinados. Uma regra que trate o iText como software de edição penaliza justamente os documentos autênticos.

O problema que orienta este relato é, portanto: até onde os metadados de arquivos PDF sustentam a detecção de indícios de adulteração em documentos nativos digitais apresentados em lote, quando a falsificação pode ter sido produzida por fabricação integral?

O objetivo geral é examinar, em um caso real com conclusão técnica conhecida, o alcance e os limites da análise de metadados na detecção desses indícios. Os objetivos específicos são:

a) descrever o desempenho de uma ferramenta que analisa cada arquivo isoladamente, aplicada a um lote de 21 documentos com parecer técnico como padrão de referência;

b) identificar, por estudo de ablação, de que componente da análise veio a separação entre documentos legítimos e não conformes;

c) discutir as fontes de falso positivo e os vetores de evasão que delimitam o uso pericial do método.

A justificativa é de ordem prática e metodológica. Na prática, a triagem automatizada de lotes documentais reduz o tempo de exame, mas uma regra mal calibrada produz erro sistemático com aparência de fundamentação técnica, como se verá na seção 4. Esse risco é

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

parte de uma discussão mais ampla: as ciências forenses têm sido cobradas a demonstrar empiricamente a validade e as taxas de erro de seus métodos, em vez de apoiá-los apenas na experiência do examinador (Saks; Koehler, 2005). Metodologicamente, conjuntos reais de comprovantes com conclusão técnica raramente estão disponíveis para pesquisa, por sigilo, e o caso aqui relatado permite observar o comportamento do método sobre documentos produzidos fora do laboratório.

2. REFERENCIAL TEÓRICO

2.1. Estrutura do PDF relevante ao exame

O formato PDF, especificado na ISO 32000-1 (International Organization for Standardization, 2008) e, em sua versão vigente, na ISO 32000-2 (International Organization for Standardization, 2020), organiza o arquivo em objetos localizados por uma tabela de referência cruzada. Ao final do arquivo, o *trailer* e o ponteiro startxref indicam onde essa tabela se encontra, e o marcador %%EOF encerra o arquivo. A partir do PDF 1.5, a tabela pode ser gravada como fluxo de referência cruzada (*cross-reference stream*), forma adotada por muitos geradores atuais.

Dessa arquitetura decorrem dois elementos que a rotina pericial inspeciona. O primeiro é a atualização incremental: alterações podem ser acrescentadas ao final do arquivo, com nova seção de referência cruzada e novo %%EOF, preservando-se o conteúdo anterior. Múltiplos marcadores indicam que o arquivo foi gravado mais de uma vez, o que não equivale a adulteração. Há causas legítimas conhecidas: arquivos linearizados, otimizados para exibição na *web*, trazem duas seções de referência cruzada e, com frequência, dois marcadores sem edição alguma; e a assinatura digital aposta a um documento já emitido é, em regra, gravada como atualização incremental. A ferramenta descrita na seção 3 desconta o marcador adicional dos arquivos linearizados.

O segundo elemento são os metadados. O dicionário /Info registra as datas de criação e de modificação e identifica o software nos campos /Creator e /Producer. Desde o PDF 1.4, o arquivo pode conter também um pacote XMP, segunda camada de metadados em XML; a divergência entre as duas camadas pode constituir indício, a ponderar conforme o gerador, pois há programas que atualizam só uma delas. No PDF 2.0, o dicionário /Info foi declarado obsoleto, com exceção das datas de criação e de modificação, o que afeta a validade futura de critérios baseados em /Creator e /Producer. Esses campos, assim como /Author, costumam carregar informação sobre quem produziu o arquivo e em que ambiente: em quase 40 mil arquivos PDF publicados por agências de segurança, Adhatarao e Lauradoux (2021) recuperaram nomes de autor, de software e de sistema operacional na maior parte dos documentos.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Desses dois elementos derivam os três indícios que a análise por arquivo examina e que a primeira versão da ferramenta descrita na seção 3 implementa: múltiplos %%EOF, derivado do primeiro; divergência entre as datas de criação e de modificação e presença de software de edição em /Producer, ambos derivados do segundo.

2.2. Edição e fabricação integral

A literatura sobre adulteração de arquivos PDF assinados descreve ataques que subvertem um documento legítimo. Os *shadow attacks* escondem, no documento a ser assinado, conteúdo que é revelado depois da assinatura sem invalidá-la (Mainka; Mladenov; Rohlmann, 2021). O *Evil Annotation Attack* e o *Sneaky Signature Attack* alteram o conteúdo visível de documentos certificados sem que os leitores de PDF emitam alerta (Rohlmann *et al.*, 2021). Em todos os casos, há um original autêntico sobre o qual o atacante opera.

A fabricação integral é de natureza diversa: não existe original. O documento é composto do zero, normalmente em processador de texto, e exportado uma única vez. Por isso, não apresenta atualização incremental, não exibe marcadores duplicados e traz datas de criação e de modificação coerentes entre si. A distinção aproxima-se da que a lei penal faz entre falsificar o documento no todo e alterar documento verdadeiro (ver Introdução). Os três indícios da seção 2.1 pressupõem uma operação de edição registrada no arquivo, que na fabricação integral não existe. Sob o critério estrutural, o documento fabricado dessa forma é indistinguível de um documento legítimo gerado do mesmo modo. O que o denuncia, quando algo o denuncia, é a incompatibilidade entre a ferramenta que o produziu e a origem que ele alega ter.

A estrutura do arquivo, contudo, registra apenas o modo do último salvamento. Um documento verdadeiro convertido para processador de texto, alterado e exportado de novo apresenta a mesma assinatura estrutural de exportação única. Os metadados não distinguem, portanto, a fabricação integral da conversão e reexportação de documento preexistente; ambas as hipóteses permanecem abertas sempre que se fala, neste relato, em documento compatível com fabricação integral.

2.3. Atribuição da ferramenta criadora

Olivier (2026) transpõe para o domínio digital o raciocínio da documentoscopia e da balística: assim como um instrumento deixa marca característica no suporte, um gerador de PDF deixa vestígios no arquivo que produz. As decisões que cada ferramenta toma livremente na montagem do arquivo, nos pontos em que a especificação admite mais de uma solução válida, funcionam como *toolmarks* estáveis, cuja ocorrência consistente o autor verifica em um conjunto de mil documentos PDF de domínios governamentais (.gov) disponibilizado pela Biblioteca do

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Congresso dos Estados Unidos. Por via distinta, Zia e Adedayo (2025) identificam o tipo de ferramenta produtora sem recorrer aos metadados, a partir de histogramas de bytes e medidas de entropia do conteúdo do arquivo, classificados por aprendizado de máquina.

Esses trabalhos estabelecem que a ferramenta produtora é tecnicamente identificável, pressuposto de que este relato depende. A pergunta que formulam é qual ferramenta produziu um documento tomado isoladamente. A pergunta aqui examinada é se a ferramenta observada é compatível com o emissor que o documento alega ter e com os demais documentos apresentados com ele. Trabalho recente, ainda sem revisão por pares, parte de premissa próxima: o mesmo produtor é normal para um emissor e anômalo para outro, e por isso a avaliação combina o emissor alegado com uma impressão digital de referência, aprendida de documentos autênticos desse emissor (El Jakani *et al.*, 2026). Essa pergunta tem raiz na documentoscopia clássica, cujo exame é comparativo por natureza: o documento questionado é confrontado com padrões de origem conhecida (Del Picchia Filho; Del Picchia; Del Picchia, 2016). O que se propõe é aplicar esse confronto aos metadados. Uma busca exploratória, não sistemática, feita nas bases Crossref e OpenAlex em novembro de 2026, limitada a publicações desde 2015 e com exame dos primeiros resultados de cada consulta, usou cinco expressões em inglês que combinavam PDF, metadados, ferramenta produtora, fraude documental e comparação entre documentos de um conjunto. Ela não localizou trabalho que tome como unidade de análise forense de PDF o conjunto de documentos apresentados juntos.

2.4. A cadeia de emissão dos sistemas institucionais

A biblioteca de relatórios JasperReports, amplamente empregada em sistemas corporativos e governamentais, gera PDF por meio da versão 2.1.7 da biblioteca iText, como declara seu descritor de dependências (Jaspersoft, [s. d.]a). As versões posteriores do iText, a partir da 5, passaram a ser distribuídas sob licença AGPL, e os mantenedores do JasperReports explicam que não podem depender delas sem adotar a mesma licença (Jaspersoft, 2017); o uso de versões recentes do iText exige extensão à parte (Jaspersoft, [s. d.]b). Por isso, documentos emitidos por esses sistemas trazem, nos metadados, o par JasperReports/iText, e sua presença é o esperado em documento autêntico emitido por eles.

Essa regularidade tem validade temporal limitada. Desde a versão 7.0.8, o JasperReports pode operar com a biblioteca OpenPDF (Jaspersoft, 2026), e há pedido registrado junto ao fabricante, desde fevereiro de 2024, para permitir a configuração do campo /Producer (Jaspersoft, 2024). Se adotadas, essas mudanças alteram a assinatura técnica dos documentos institucionais, e qualquer regra baseada nela precisará ser revista.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3. METODOLOGIA

O trabalho segue a *Design Science Research*, em que a construção e a avaliação de um artefato constituem a própria pesquisa (Hevner *et al.*, 2004). O artefato, uma instanciação de software, é a ferramenta descrita nas seções 3.3 a 3.5. Das seis etapas propostas por Peffers *et al.* (2007), o relato cobre a definição de objetivos, o projeto e desenvolvimento, a demonstração e a avaliação, e este artigo constitui a etapa de comunicação. Foram realizados dois ciclos de construção e avaliação: a primeira versão falhou na avaliação, e a análise dessa falha orientou a segunda. Os critérios de avaliação foram três: ordenar os documentos não conformes acima dos legítimos, medido pela AUC (seção 3.6); acompanhar cada indício da evidência bruta que o sustenta; e não penalizar a cadeia de emissão institucional legítima. A avaliação é observacional, do tipo estudo de caso, feita sobre o mesmo material que orientou a construção da segunda versão; por isso descreve o comportamento do artefato nesse caso e não o confirma para casos novos.

3.1. Material

O material consiste em 21 documentos nativos digitais em formato PDF, apresentados conjuntamente, em um mesmo ato, e submetidos a exame técnico de documento digital já concluído. Neste trabalho, lote designa o conjunto de documentos que chega ao exame como unidade, por ter sido apresentado em conjunto. O mesmo exame abrangeu também imagens apresentadas como capturas de tela, excluídas deste estudo por não serem documentos nativos digitais em PDF (seção 4.9). A definição não pressupõe emissor único, e o lote examinado não o tem. Ele reúne cinco grupos:

- guias de recolhimento emitidas por sistema de órgão público (D01–D03);
- comprovantes de recolhimento referentes aos mesmos períodos (D04–D06);
- comprovantes de parcelas de uma obrigação periódica (D07–D10);
- boleto de origem digitalizada (D11);
- comprovantes bancários de um mesmo emissor (D12–D21).

Os dois grupos de documentos legítimos provêm de emissores distintos, mas compartilham a mesma família de ferramentas de geração (JasperReports/iText), o que permite estabelecer uma cadeia dominante no lote (seção 3.4). Não há, no lote, comprovante de parcela legítimo. Por isso, o perfil esperado para essa tipologia não foi obtido de pares, e sim de uma tabela fixa que associa cada tipo de documento ao tipo de sistema que deveria emiti-lo.

O padrão de referência são as conclusões do parecer técnico emitido no caso por perito de escritório privado e juntado ao processo, que não é laudo de perito oficial (Brasil, 1941, art.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

159) e é aqui chamado apenas de parecer: 13 documentos legítimos, 7 não conformes e 1 inconclusivo. O documento inconclusivo (D11) foi excluído do cômputo de desempenho, que abrange, portanto, 20 documentos. Como o parecer se limitou aos metadados (ver adiante), a classe "legítimo" deve ser lida como "compatível com a emissão alegada", e não como autenticidade confirmada junto ao emissor; da mesma forma, "não conforme" designa incompatibilidade técnica, e não juízo sobre fato ainda sujeito a julgamento. O parecer é uma conclusão técnica humana, sujeita a erro, e não uma verdade construída pelo pesquisador. Isso distingue este material dos conjuntos de adulterações sintéticas, em que o rótulo é conhecido porque o próprio pesquisador produziu a falsificação, mas também impõe ao resultado o limite de confiabilidade do próprio parecer. O parecer não é independente dos critérios avaliados: seu achado central, a coincidência de data e hora de criação (seção 4.3), é um dado de metadados, e o confronto entre o software produtor e a origem alegada integra o raciocínio técnico que o produziu. Além do software produtor e das datas, o parecer considerou as fontes tipográficas embarcadas, a proximidade entre os momentos de criação e a nomenclatura dos arquivos; declara expressamente que a análise se limitou aos metadados e apenas recomenda, sem tê-la realizado, a confirmação dos pagamentos junto às instituições emissoras. A concordância entre ferramenta e parecer reproduz, portanto, parte do raciocínio que gerou o rótulo. Em estudos de desempenho diagnóstico, essa situação é conhecida como viés de incorporação, que tende a superestimar a concordância entre o teste e a referência (Worster; Carpenter, 2008), e os resultados são lidos com essa restrição (seção 4.8).

3.2. Sigilo, autorização e anonimização

O uso do material foi autorizado por escrito pelo perito que assinou o parecer, coautor deste relato. O termo de autorização delimita a finalidade acadêmica, as condições de anonimização e a não divulgação dos arquivos. Quem contratou o exame concordou com o uso acadêmico do material anonimizado, e sua identidade não é revelada. No termo, o perito declara ainda que, até onde é de seu conhecimento, não há impedimento legal à publicação anonimizada, inclusive segredo de justiça sem a devida autorização. A pesquisa enquadra-se no art. 1º, parágrafo único, VII, da Resolução CNS nº 510/2016 (Brasil, 2016), que dispensa de registro no sistema CEP/Conep a pesquisa que objetiva o aprofundamento teórico de situação surgida na prática profissional, desde que não revele dados que permitam identificar os envolvidos. A Comissão Nacional de Ética em Pesquisa esclarece que a hipótese pressupõe a ausência de ação diferente da prática cotidiana e a impossibilidade de identificar os participantes (Brasil, 2022). O tratamento, com finalidade exclusivamente acadêmica, observa a LGPD (Brasil, 2018, arts. 4º, II, b, e 7º): os dados publicados são anonimizados (art. 12), e o mapeamento entre

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

identificadores e arquivos, que constitui pseudonimização (art. 13, § 4º), permanece restrito aos autores.

A ferramenta foi executada sobre os arquivos originais, sem nenhuma alteração: os resumos SHA-256 dos 21 arquivos analisados coincidem com os registrados no parecer, procedimento de preservação da integridade que segue as diretrizes da ISO/IEC 27037 para evidência digital (International Organization for Standardization, 2012), o que assegura a identidade entre o material aqui analisado e o examinado no parecer. A anonimização incide sobre o que se publica. Os nomes de arquivo foram substituídos pelos identificadores D01–D21, e o mapeamento é mantido à parte, fora de qualquer repositório. Nenhum nome de pessoa física ou jurídica, inscrição fiscal, número de processo, endereço, valor ou campo de metadados que identifique pessoa ou equipamento é reproduzido neste artigo; a coincidência de data e hora de criação é descrita sem os valores, e os grupos de documentos são descritos de forma genérica, para reduzir o risco de reconhecimento do caso por terceiros.

Durante a pesquisa, os *scripts* de avaliação foram executados com auxílio de ferramenta de inteligência artificial generativa de uso comercial (ver Declarações), sob supervisão dos autores. Esse uso implicou o processamento, pelo serviço, dos arquivos do caso, ainda não anonimizados, de seus metadados e do texto do parecer. Nesse tratamento de dados pessoais, o fornecedor do serviço atua como operador (Brasil, 2018, art. 5º, VII). O serviço foi utilizado com a opção de uso das conversas para treinamento de modelos desativada. O perito signatário declarou, no termo de autorização, ciência desse processamento. Os autores registram o procedimento como limitação e recomendam, em pesquisas semelhantes, extrair os metadados localmente e submeter a serviços externos apenas dados já anonimizados.

Os documentos não são disponibilizados. A anonimização do conteúdo visível de um PDF não anonimiza seus metadados, e são justamente os metadados o objeto do exame: publicar os arquivos reintroduziria o risco que a anonimização pretende afastar. O código-fonte da ferramenta, na versão que produziu os resultados aqui relatados (seção 3.6), acompanha este artigo como arquivo suplementar, com manual de execução e sem os arquivos do caso; para a distribuição, os comentários explicativos do código foram retirados, sem alteração de nenhuma instrução executável.

3.3. Versão por arquivo

A primeira versão da ferramenta, escrita em Python (executada com Python 3.13.13, pypdf 6.14.2 e PyMuPDF 1.28.2), analisa cada arquivo isoladamente. Extrai do dicionário /Info as datas de criação e de modificação e os campos que identificam a aplicação criadora e a produtora; sobre o binário, conta os marcadores de fim de arquivo e as revisões, descontando o

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

marcador adicional dos arquivos linearizados, que não indica edição. Em documentos assinados, verifica a cobertura da assinatura. Cada indício soma um peso ao escore, limitado a 100 (Quadro 1). Nenhuma das duas versões lê a camada XMP.

Quadro 1. Indícios e pesos da versão por arquivo

Indício	Peso
Modificação de conteúdo após assinatura digital	45
Bytes fora da cobertura da assinatura	30
Dados anexados após o último %%EOF	30
Documento não assinado com mais de uma revisão (atualização incremental)	25
Software classificado como de edição em /Creator ou /Producer	20
/Creator e /Producer ausentes	15
Falha de leitura da estrutura	15
Data de modificação diferente da de criação	10
Certificação que permite alterações posteriores (DocMDP P=3)	5

Fonte: elaborado pelos autores.

A base de conhecimento dessa versão classificava o iText como software de edição. Essa classificação é a origem da inversão relatada na seção 4.

3.4. Versão contextual

A segunda versão não reaproveita os indícios estruturais da primeira: é uma função de escore distinta, que avalia cada documento contra o tipo que ele alega ser e contra os demais documentos do lote (Quadro 2).

Quadro 2. Critérios e pesos da versão contextual

Critério	Natureza	Regra	Peso
Perfil de emissor	por arquivo, com informação externa	Software produtor de natureza manual (processador de texto) em documento que alega origem institucional	40
		Software de outra natureza incompatível (por exemplo, digitalização)	20
		Metadados de origem ausentes ou software não reconhecido pela tabela de perfis, impedindo confirmar a emissão	20
Linha de base do lote	relacional	Família de ferramentas diferente da dominante no lote (aplicado só se a dominante cobre ao menos 30% dos documentos)	15
Data e hora de criação idênticas	relacional	Iguais, ao segundo, às de outro documento de período de referência distinto	45
		Idem, mesmo período	20
Defasagem temporal	por arquivo, com informação externa	Criação 3 meses ou mais após o período de referência	30
		Criação 2 meses após o período (tolerância de 1 mês)	15
Modificação posterior	por arquivo	Modificação mais de 7 dias após a criação	20
		Modificação entre 1 e 7 dias após a criação	10

Fonte: elaborado pelos autores. Nota: o escore é a soma dos pesos, limitada a 100.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Alguns detalhes de implementação condicionam a leitura dos resultados. A linha de base compara o par de famílias de ferramenta (criadora e produtora), e não a versão, para que versões diferentes do mesmo sistema não sejam tratadas como divergentes; documento sem metadados de origem não tem família e não é comparado, mas entra no total sobre o qual se calcula a proporção de 30%. A tabela fixa de perfis associa todas as tipologias do lote à emissão automatizada e classifica cada família de software em três naturezas: emissão automatizada, produção manual e digitalização; software ausente ou não reconhecido pela tabela recebe o peso de 20 pontos atribuído à falta de metadados de origem. A comparação de datas e horas de criação exige igualdade exata ao segundo, desconsidera o deslocamento de fuso horário e só distingue períodos quando o mês consta do nome dos dois arquivos; a sequência de navegação considera intervalos de 5 segundos a 5 minutos entre ao menos três documentos e não pontua. O tipo de documento e o período de referência são lidos do nome com que o arquivo foi recebido para exame, e não do conteúdo; quando o ano não consta do nome, é inferido a partir da data de criação. Como a tabela associa todas as tipologias do lote à emissão automatizada, o tipo é constante neste lote: funciona como a premissa de que o documento alega origem institucional, e não distingue um documento de outro. A linha de base e a modificação posterior não dependem do nome, e a comparação de data e hora de criação o usa apenas para distinguir períodos. Os nomes foram atribuídos por quem encaminhou os documentos ao exame, antes das conclusões do parecer e independentemente delas, e constam de seu registro de integridade; a padronização feita depois, na anonimização, não alterou nenhum resultado (seção 4.2).

3.5. Definição dos pesos e do escore

Os pesos dos Quadros 1 e 2 não foram estimados estatisticamente, o que exigiria um conjunto de documentos rotulados muito maior que o disponível. Foram atribuídos por juízo pericial, sem uma escala formal registrada no projeto. Examinados em conjunto, os valores se organizam em quatro níveis ordinais, que podem ser lidos à luz de duas perguntas: quanto o indício, sozinho, afasta o documento do que se espera de uma emissão legítima do tipo alegado, e com que frequência ele admite explicação inocente. Os indícios de 40 a 45 pontos são os de explicação inocente rara, como o processador de texto em documento que pressupõe emissão automatizada; os de 25 a 30, de explicação ocasional; os de 15 a 20, de explicação frequente ou que apenas impedem confirmar a origem; e os de 5 a 10, de explicação muito frequente. Essa organização descreve os valores atribuídos e não foi usada para defini-los. A atualização incremental ilustra essa leitura. Sem assinatura digital, ela indica que o arquivo foi gravado de novo depois de emitido e vale 25 pontos; quando a atualização é a própria assinatura, não

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

pontua, porque a assinatura aposta a documento já emitido é, em regra, gravada desse modo; quando modifica o conteúdo depois da assinatura, vale 45; quando acrescenta bytes fora da assinatura sem conteúdo classificável, vale 30.

O escore de cada documento é a soma dos pesos dos indícios acionados, limitada a 100. Por exemplo, D04 acionou perfil de emissor (40), divergência da linha de base (15), data e hora de criação idênticas às de documento de outro período (45) e defasagem de três meses ou mais (30): a soma, 130, é limitada a 100; D06 acionou um único indício, metadados ausentes (20); os legítimos não acionaram nenhum (Tabela 1). A soma supõe que os indícios se acumulam de forma independente, o que não é exato: o perfil de emissor e a linha de base podem apontar o mesmo processador de texto e contá-lo duas vezes. Por isso o escore deve ser lido como ordinal. Ele ordena os documentos para a triagem, mas não é probabilidade de falsificação nem razão de verossimilhança, que é a forma recomendada na literatura para expressar o valor probatório de um achado (Aitken; Taroni, 2004; European Network of Forensic Science Institutes, 2015). Pela mesma razão, o desempenho é avaliado por medidas que dependem só dessa ordenação (seção 3.6).

O limiar de 20 foi fixado na etapa de avaliação, já conhecido o resultado do parecer, e não decorre das faixas de classificação que a ferramenta usa em seu relatório. Por isso não é tomado como critério de desempenho. Na escala de níveis descrita acima, ele equivale a encaminhar ao exame humano qualquer documento que acione ao menos um indício de 20 pontos ou mais; indícios de 15 pontos ou menos só o alcançam somados. Na versão contextual, como nenhum documento legítimo acionou indício, qualquer conjunto de pesos positivos produziria a mesma separação (AUC de 1,00 e margem positiva): os valores escolhidos afetam a magnitude dos escores, a ordem entre os não conformes e a posição de cada um em relação ao limiar, mas não a separação entre legítimos e não conformes. O que depende do conhecimento prévio do caso são os próprios critérios, e não seus pesos (seção 3.7).

3.6. Ablação, métricas e limiar

Para separar a contribuição de cada componente, cada versão foi executada uma única vez sobre o lote, e os escores de sete configurações foram obtidos somando apenas os indícios correspondentes, sem alteração de pesos:

- C1 — versão por arquivo, original;
- C2 — versão por arquivo, sem o indício atribuído ao iText como software de edição;
- C3 — versão contextual, sem os critérios relacionais (linha de base e data e hora de criação idênticas);
- C4 — versão contextual completa;

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

• C5 — versão contextual, apenas com o critério de perfil de emissor;

• C6 — versão contextual, apenas com os critérios relacionais;

• C7 — versão contextual, apenas com a linha de base do lote.

C7 é a única configuração que não usa nenhuma informação do nome do arquivo; C6 dispensa o tipo, mas usa o período lido do nome. Para medir o papel do nome, a versão contextual foi reexecutada sobre os mesmos metadados em três condições: com os nomes registrados no parecer, com o mesmo tipo atribuído a todos os documentos e com nomes neutros, sem tipo reconhecível.

O desempenho é descrito por métrica independente de limiar, a área sob a curva ROC (AUC), que mede a capacidade do escore de ordenar os casos positivos acima dos negativos (Fawcett, 2006; Hanley; McNeil, 1982). Ela foi calculada como a proporção de pares (não conforme, legítimo) em que o não conforme pontua acima, com empates contados como meio. Usa-se também a margem, definida como a diferença entre o menor escore dos não conformes e o maior escore dos legítimos: margem positiva indica separação completa. Para o limiar de 20 (seção 3.5), aplicado igualmente a todas as configurações, reportam-se a acurácia (proporção de documentos corretamente classificados), a revocação e a especificidade, estas em valores absolutos, com intervalos de confiança de 95% pelo método de Wilson (Wilson, 1927), recomendado para proporções em amostras pequenas (Brown; Cai; Dasgupta, 2001), além da faixa de limiares que separa completamente o lote. Como o método supõe observações independentes, o que não ocorre aqui (seção 4.8), os intervalos são apenas ilustrativos. As fontes de falso positivo e os vetores de evasão (objetivo c) foram derivados analiticamente das regras dos Quadros 1 e 2 e da especificação do formato, e não testados no lote. As regras e os pesos são os da versão do código identificada como v3.0-avaliada-L1 (*commit* 5337499), fornecida como arquivo suplementar; as configurações C6 e C7 e as reexecuções com outros nomes foram acrescentadas depois aos *scripts* de avaliação, sem alteração de regras ou de pesos.

3.7. Posição dos autores em relação ao caso

Dois dos autores participaram do exame técnico que originou o caso: um deles o conduziu, sob supervisão do outro, perito responsável que assinou o parecer usado como padrão de referência. As conclusões do parecer eram, portanto, conhecidas quando a versão contextual foi construída, e os pesos do Quadro 2 foram calibrados neste mesmo lote. Não houve cegamento. A literatura sobre fatores humanos na perícia mostra que o conhecimento prévio da conclusão tende a orientar a busca e a interpretação dos indícios na direção dela (Dror, 2020; Kassin; Dror; Kukucka, 2013). Os resultados descrevem, portanto, a consistência dos critérios com um caso conhecido e não estimam desempenho em casos novos.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

4. RESULTADOS E DISCUSSÃO

4.1. Escores por documento

A Tabela 1 apresenta, para cada documento, o escore e os indícios acionados nas duas versões da ferramenta. Os treze documentos classificados como legítimos no parecer obtiveram escores e indícios idênticos entre si em cada versão e aparecem agrupados.

Tabela 1. Escores e indícios por documento nas duas versões

Doc.	Tipologia	Parecer	Por arquivo	Indícios	Contextual	Indícios
D01–D03	Guias (3)	Legítimo	20	SE 20	0	—
D04	Recolhimento	Não conforme	0	—	100	PE 40; LB 15; CI 45; DT 30
D05	Recolhimento	Não conforme	0	—	100	PE 40; LB 15; CI 45; DT 30
D06	Recolhimento	Não conforme	15	MA 15	20	MA 20
D07	Parcela	Não conforme	25	AI 25	55	PE 40; LB 15
D08	Parcela	Não conforme	10	DM 10	90	PE 40; LB 15; DT 15; MP 20
D09	Parcela	Não conforme	25	AI 25	55	PE 40; LB 15
D10	Parcela	Não conforme	25	AI 25	70	PE 40; LB 15; DT 15
D11	Boleto digitalizado	Inconclusivo	0	—	35	PE 20; LB 15
D12–D21	Pagamento (10)	Legítimo	20	SE 20	0	—

Fonte: elaborada pelos autores.
Nota: o escore é a soma dos pesos indicados, limitada a 100. Recolhimento: comprovante de recolhimento; Parcela: comprovante de parcela; Guias: guias de recolhimento; Pagamento: comprovantes bancários. AI: atualização incremental; CI: data e hora de criação idênticas às de documento de período distinto; DM: data de modificação diferente da de criação; DT: defasagem temporal (30 se ≥ 3 meses, 15 se de 2 meses); LB: divergência da linha de base do lote; MA: metadados de origem ausentes ou não reconhecidos; MP: modificação mais de 7 dias após a criação; PE: perfil de emissor incompatível; SE: iText classificado como software de edição.

Na versão por arquivo, os legítimos tiveram escore médio de 20,00 e os não conformes, de 14,29. Os treze legítimos pontuaram exatamente 20, todos pelo mesmo motivo: a presença do iText, classificado como software de edição. Entre os não conformes, três (D07, D09, D10) acionaram o indício de atualização incremental, isto é, apresentam mais de um marcador %%EOF; o conteúdo das revisões não foi examinado. É o único indício estrutural observado no lote. Os dois documentos de escore zero (D04 e D05) não apresentaram nenhuma

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

anomalia no arquivo. Na versão contextual, as médias foram 0,00 para os legítimos e 70,00 para os não conformes.

4.2. Estudo de ablação

A Tabela 2 apresenta o desempenho das sete configurações, segundo as medidas definidas na seção 3.6; na AUC, 0,5 corresponde a ordenação ao acaso e 1,0 a separação perfeita.

Tabela 2. Desempenho por configuração (n = 20; 7 não conformes, 13 legítimos)

Configuração	AUC	Margem	Revocação, limiar 20 (IC 95%)	Especificidade, limiar 20 (IC 95%)	Limiares que separam todo o lote
C1 — por arquivo, original	0,43	−20	3/7 (0,16–0,75)	0/13 (0,00–0,23)	nenhum
C2 — por arquivo, sem penalizar o iText	0,86	0	3/7 (0,16–0,75)	13/13 (0,77–1,00)	nenhum
C3 — contextual, sem critérios relacionais	1,00	+20	7/7 (0,65–1,00)	13/13 (0,77–1,00)	1 a 20
C4 — contextual, completa	1,00	+20	7/7 (0,65–1,00)	13/13 (0,77–1,00)	1 a 20
C5 — contextual, só perfil de emissor	1,00	+20	7/7 (0,65–1,00)	13/13 (0,77–1,00)	1 a 20
C6 — contextual, só critérios relacionais	0,93	0	2/7 (0,08–0,64)	13/13 (0,77–1,00)	nenhum
C7 — contextual, só linha de base do lote	0,93	0	0/7 (0,00–0,35)	13/13 (0,77–1,00)	nenhum

Fonte: elaborada pelos autores.
Nota: intervalos de confiança de Wilson, ilustrativos (seção 3.6). Critérios relacionais: linha de base do lote e data e hora de criação idênticas. Em C5, a regra de perfil de emissor inclui o caso de metadados de origem ausentes. Em C7, o peso único da regra (15) fica abaixo do limiar de 20.

Na versão por arquivo, a média dos legítimos superou a dos não conformes, mas a ordenação ficou próxima do acaso (AUC de 0,43). Com limiar 21, C1 teria acurácia de 0,80; a acurácia de 0,15 no limiar 20 resulta em boa parte de o limiar coincidir com o escore de todos os legítimos. Retirada a penalização do iText (C2), a AUC passa a 0,86 e os falsos positivos desaparecem, mas D04 e D05 permanecem com escore zero, igual ao dos legítimos. Nenhum limiar separa o lote nessa configuração; a melhor acurácia, 0,90, ocorre entre os limiares 1 e 10.

A regra de perfil de emissor separa sozinha o lote (C5), em qualquer limiar de 1 a 20: seis não conformes pelo processador de texto e D06 pela ausência de metadados de origem. A linha de base, também sozinha (C7), atribui 15 pontos a seis dos sete não conformes e zero a todos os legítimos: entre os limiares 1 e 15, acerta 19 dos 20 documentos (acurácia de 0,95),

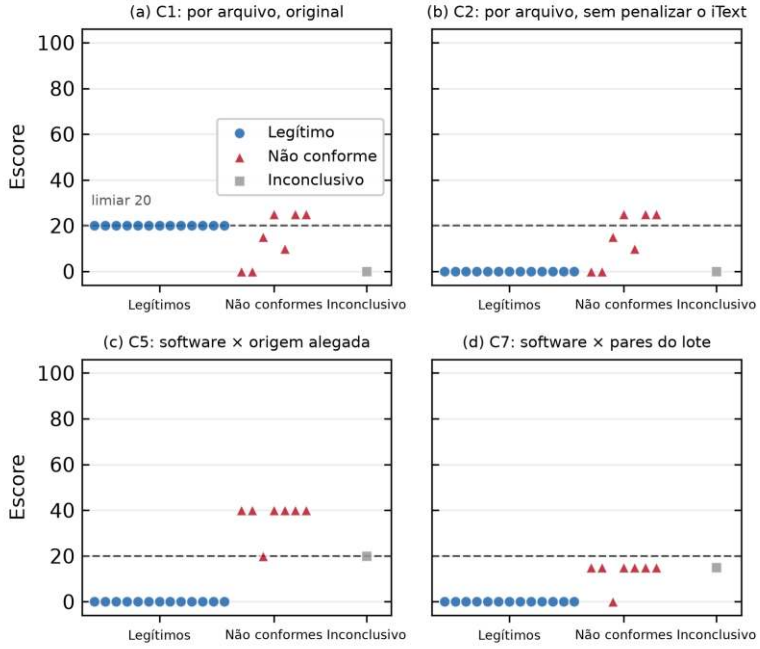
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

mas no limiar de 20, acima de seu único peso, não alcança nenhum (0/7). O único não conforme que ela não alcança é D06, que não tem metadados de origem e, por isso, não tem família a comparar. Os dois critérios relacionais juntos (C6) mantêm a AUC de 0,93, toda ela proveniente da linha de base, e elevam D04 e D05 a 60 pontos, pela coincidência de data e hora de criação.

Reexecutada com os nomes do parecer ou com o mesmo tipo atribuído a todos os documentos, C5 não muda (AUC de 1,00): o conteúdo do nome não discrimina. Com nomes neutros, a regra não encontra tipo reconhecível e não dispara: C5 cai a zero em todos os documentos (AUC de 0,50), C7 permanece igual e a versão completa (C4) fica com AUC de 0,93 e 3/7 no limiar de 20. O nome funciona, portanto, como um interruptor: a regra exige a premissa de origem institucional, e não a distinção entre tipos.

A Figura 1 mostra a dispersão dos escores nessas quatro configurações (C1, C2, C5 e C7).

Figura 1. Dispersão dos escores nas configurações C1, C2, C5 e C7, por classe do parecer



Fonte: elaborada pelos autores.

Somados aos critérios por arquivo, os relacionais não alteram nenhuma classificação nos limiares de 1 a 20: C3 e C4 têm a mesma AUC e a mesma faixa de separação. Como C3 já atinge AUC de 1,00, a ablação não tem como medir contribuição adicional acima desse teto.

4.3. Achados de conjunto

Os critérios relacionais produziram dois registros que a análise por arquivo não alcança.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Coincidência de data e hora de criação. D04 e D05, referentes a períodos distintos, registram em /Info data e hora de criação idênticas ao segundo; esse valor é declarado pelo programa gerador e não tem a garantia de um carimbo do tempo. É o achado central do parecer. O critério, uma vez formalizado, recupera-o automaticamente. A recuperação não é descoberta independente, pois o critério foi concebido depois de conhecido o parecer (seção 4.8).

Sequência compatível com navegação manual. Os dez comprovantes bancários legítimos (D12–D21) foram criados em sequência, com intervalos de 13 a 64 segundos entre um e outro. O padrão é compatível com um usuário que solicita os documentos um a um em portal *web*, e também o seria com a produção manual de arquivos em sequência. É elemento compatível com a autenticidade, sem peso probatório próprio.

O parecer apontou ainda um terceiro registro de conjunto que a ferramenta não alcança: dois comprovantes de parcela (D07 e D09), referentes a meses distintos, foram criados no mesmo dia, com poucos minutos de diferença, o que o parecer interpretou como produção em lote. O critério da ferramenta exige igualdade ao segundo e não registra essa proximidade. Uma janela temporal mais larga a alcançaria, ao preço de maior risco de falso positivo na emissão legítima em bloco (seção 4.7).

A linha de base do lote foi identificada por frequência, sem parametrização manual: a cadeia JasperReports/iText está presente em 13 dos 21 documentos. Dos oito restantes, seis apresentam metadados de processador de texto, um de aplicativo de digitalização (D11) e um não tem metadados de origem (D06).

4.4. Casos-limite

D06. Na versão contextual, pontuou exatamente 20, sustentado por um único indício de natureza negativa: a ausência de metadados de origem, que impede confirmar a cadeia de emissão. Qualquer limiar entre 1 e 20 separa o lote; com limiar 21, D06 passaria a falso negativo e a revocação cairia para 6/7 (0,86). É também o único não conforme que a linha de base não alcança. Foi criado no próprio mês de referência, de modo que nem a defasagem nem a coincidência de data e hora se aplicam; no parecer, sua classificação apoiou-se na ausência de identificação do emissor e nas fontes tipográficas, das quais a ferramenta reproduz apenas a primeira. A mesma ausência de metadados ocorre em documentos legítimos reemitidos ou rasterizados, o que torna esse indício o ponto mais frágil do resultado.

D11. O boleto de origem digitalizada obteve escore 35 e ficou fora do cômputo, por ter sido classificado como inconclusivo no parecer. A digitalização suprime a cadeia de emissão original por construção, e não por adulteração, de modo que os indícios acusados admitem

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

explicação inocente. D11 pontuou acima de D06, documento não conforme: o escore mede afastamento do padrão de emissão esperado, e não probabilidade de falsificação.

4.5. Por que a análise por arquivo falhou

A inversão da versão por arquivo não decorreu de erro de implementação: a ferramenta executou o que fora projetada para executar. O ponto incorreto era a base de conhecimento, que tratava o iText como software de edição. Como os treze documentos legítimos foram emitidos por sistemas que usam essa biblioteca (seção 2.4), todos receberam a mesma penalização, e a ferramenta passou a apontar como suspeito exatamente o padrão da emissão legítima. O erro é plausível, porque o iText também serve para modificar arquivos PDF existentes, mas parece próprio desta ferramenta: em consulta não exaustiva à documentação pública, em novembro de 2026, não se localizou ferramenta de triagem que o classifique como software de edição. A título apenas ilustrativo, um verificador comercial de extratos o lista entre as assinaturas esperadas de emissão bancária (Terra Insight, 2026).

Corrigida a base de conhecimento (C2), a análise por arquivo deixa de errar sobre os legítimos e continua sem alcançar dois dos sete documentos não conformes. A Tabela 1 mostra que os não conformes não formam um grupo homogêneo. Em três comprovantes de parcela (D07, D09 e D10), o binário apresenta dois marcadores de fim de arquivo e dois ponteiros startxref, sem linearização, e o indício de atualização incremental os alcançou. A data de modificação desses arquivos, contudo, é igual à de criação: se houve regravação, ela não atualizou o campo, o que reforça que as datas internas tampouco são confiáveis isoladamente. D04 e D05, ao contrário, não apresentam nenhuma anomalia no arquivo: têm a assinatura estrutural de exportação única, compatível com fabricação integral ou com reexportação de documento preexistente (seção 2.2), e é justamente sobre eles que a análise por arquivo, mesmo corrigida, fracassou neste lote.

4.6. O que separou o lote

D04 e D05 trazem, nos metadados, um processador de texto como software criador. O dado estava no arquivo desde o início, e ainda assim a análise por arquivo não o tratou como indício. A razão é que um processador de texto não é, por si, sinal de falsificação: é a ferramenta legítima de produção de incontáveis documentos, e uma lista de softwares suspeitos que o incluísse acusaria todos eles. O metadado informa a procedência do arquivo; para que essa procedência se torne indício, é preciso saber de onde o documento deveria ter vindo.

No caso examinado, essa referência veio de duas fontes externas ao arquivo: a primeira, sozinha, separou todo o lote; a segunda alcançou seis dos sete não conformes, nos limiares de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

1 a 15. A primeira é a origem institucional alegada (C5): um comprovante produzido por processador de texto, ou sem metadados de origem, é incompatível com a emissão automatizada que ele alega ter. Não foi a distinção entre tipos de documento que separou o lote, pois todos pressupõem essa emissão (seção 4.2). A segunda é a cadeia dominante entre os pares (C7): o processador de texto diverge da cadeia de ferramentas que predomina entre os documentos apresentados em conjunto. O lote não permite dizer qual das duas é a decisiva, porque nele ambas apontam para os mesmos documentos, com exceção de D06. Mais do que isso: como todas as tipologias do lote pressupõem emissão automatizada e nenhum documento legítimo foi produzido em processador de texto, as duas referências coincidem, neste lote, com uma regra simples por arquivo, a de considerar suspeito todo software produtor diferente de JasperReports/iText. Os dados não distinguem, portanto, o confronto com uma referência externa dessa regra simples; a vantagem do confronto é argumento conceitual, a testar em lotes com documentos legítimos de origens variadas.

É nesse sentido que os resultados sustentam a tese do relato. Os metadados de um arquivo, lidos isoladamente, indicam a procedência do arquivo, isto é, com que ferramenta foi produzido. Não bastam para afirmar se o documento é autêntico: o que torna o software produtor um indício é a origem que o documento alega ter, e essa referência o arquivo não contém. Pode-se objetar que o nome do arquivo também é um dado sobre ele. A diferença é de natureza: o software produtor é registrado pela ferramenta que gerou o arquivo, ao passo que o nome é uma declaração de quem o apresenta, que pode ser escolhida livremente. O que separou o lote foi o confronto entre o registro técnico e a origem alegada.

As duas referências têm fragilidades distintas. A origem alegada foi inferida do nome dado ao arquivo por quem o encaminhou ao exame, elemento que pode ser escolhido livremente: com nome neutro, a implementação avaliada não aplica a regra. O perito, porém, costuma saber por outra via o que o documento alega ser, e a regra poderia receber essa informação diretamente. A linha de base dispensa o nome, e C7 mostra que seis dos sete não conformes receberiam escore positivo com nomes neutros. Ela depende, em troca, de que os documentos legítimos sejam maioria no lote, condição discutida na seção 4.7.

Quanto aos critérios relacionais, a expectativa que motivou a versão contextual era a de que a evidência discriminante estaria na relação entre os documentos. Neste lote, esses critérios foram redundantes com o perfil de emissor, e não irrelevantes: sozinhos, deram escore positivo a seis dos sete não conformes e a nenhum legítimo, sem usar o tipo de documento. A coincidência de data e hora de criação entre D04 e D05 recuperou o achado central do parecer, mas os dois documentos já estavam separados pelo perfil de emissor. A hipótese de que a análise relacional se torna necessária quando a fabricação é mais sofisticada, isto é, quando

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

quem falsifica usa a ferramenta esperada para o tipo de documento, permanece plausível e não foi testada aqui. Ela tem, ademais, um contra-argumento evidente: quem conhece o método pode gerar cada arquivo em momento distinto e com a cadeia de ferramentas esperada, neutralizando tanto o perfil de emissor quanto a correlação temporal.

4.7. Fontes de falso positivo

Cinco situações legítimas, nenhuma representada no lote, produziriam indícios sobre documento autêntico. A primeira é a emissão em bloco: portais que emitem de uma só vez guias de vários períodos geram datas e horas de criação idênticas sem falsificação, e o critério não distingue esse caso da geração manual em sessão única. A segunda é a emissão de segunda via, cuja data de criação é muito posterior ao período que documenta e que o critério de defasagem sinalizaria; a distinção depende do conteúdo, que pode declarar a reemissão. A terceira é a ausência de metadados em documentos reemitidos por aplicativos ou rasterizados no caminho até o processo, o mesmo indício que sustentou, sozinho, a detecção de D06; regra análoga atinge o emissor legítimo cujo programa gerador não conste da tabela de perfis, que recebe os mesmos 20 pontos. A quarta é a reimpressão: comprovante autêntico reimpresso em PDF por impressora virtual de processador de texto ou de sistema operacional, antes de ser juntado aos autos, passa a exibir software de produção manual e receberia 40 pontos pelo perfil de emissor. A quinta é a composição do lote. A linha de base é a família de ferramentas mais frequente no conjunto. Se os documentos falsificados forem maioria, a linha de base passa a ser a deles, e os documentos autênticos é que divergem dela. A ferramenta só aplica o critério quando a família dominante cobre ao menos 30% dos documentos, valor fixado sem estudo de sensibilidade e que não impede essa captura. No lote examinado, a família dominante era a dos documentos legítimos (13 de 21), mas essa condição não pode ser presumida.

4.8. Limitações

O resultado tem valor de estudo de caso, e não de validação. Trata-se de lote único, com vinte documentos no cômputo de desempenho, que tampouco são observações independentes: os treze legítimos compartilham a mesma cadeia de ferramentas, D04 e D05 foram criados no mesmo segundo e D07 a D10 têm o mesmo software produtor. O número efetivo de padrões distintos é bem menor que vinte, e nenhuma inferência estatística se sustenta. Os critérios da versão contextual foram concebidos com o resultado conhecido, por quem participou do exame, sem cegamento (seção 3.7), e o padrão de referência está sujeito à mesma limitação que este relato demonstra: apoiou-se em metadados e não confirmou os pagamentos junto aos emissores (seção 3.1). A concordância com ele não demonstra autenticidade nem falsidade dos

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

documentos, e a AUC de 1,00 é desempenho aparente. Há informação, recebida no curso do caso e não verificada pelos autores, de que laudo oficial produzido no processo apontou os mesmos achados; como não foi possível consultá-lo, ele não é usado aqui como referência. O tipo e o período foram lidos do nome do arquivo, e não do conteúdo; a defasagem negativa, de documento criado antes do período que documenta, não gera indício; e a ferramenta lê apenas o dicionário /Info, sem examinar a camada XMP e o histórico de edição que ela pode registrar. A desconsideração do deslocamento de fuso horário pode tanto gerar coincidência aparente de data e hora entre arquivos gerados em fusos distintos quanto deixar de detectá-la; essa regra, o limiar de 30% da linha de base e as faixas de defasagem foram fixados sem análise de sensibilidade. Por fim, as duas versões diferem em mais de um aspecto ao mesmo tempo; a ablação isola os componentes dentro deste lote, mas não substitui a avaliação em conjuntos independentes.

4.9. Vetores de evasão

O alcance do método é delimitado por procedimentos que ele não detecta, e cuja não detecção se declara como resultado. A reescrita integral do arquivo por editor comercial elimina resíduos de atualização incremental e torna a cadeia de ferramentas observada coerente com a do editor. A alteração direta dos metadados por utilitário de linha de comando forja simultaneamente o perfil de emissor e as datas. A geração do documento falso com a mesma biblioteca usada pelo emissor, como o próprio iText, produz arquivo com a assinatura técnica esperada. A alteração da página de um portal bancário no próprio navegador, seguida de "salvar como PDF", gera arquivo com o produtor do navegador, que a tabela de perfis classifica como emissão automatizada. A edição por atualização incremental seguida de assinatura digital com qualquer certificado faz a atualização parecer parte da assinatura e elimina o indício correspondente. A geração dos arquivos em sessões separadas neutraliza a correlação temporal. A predominância de documentos falsificados no lote captura a linha de base. E a impressão seguida de digitalização e reconhecimento óptico destrói a estrutura original, produzindo documento novo, sem herança do anterior.

Decorre daí a delimitação do escopo: o método se aplica ao documento original nativo digital. Captura de tela e documento digitalizado estão fora de alcance por construção, pois uma imagem não possui as estruturas que o exame interroga. Nesses casos, o escore não mede falsificação: a ausência de indício estrutural decorre da falta de substrato, e o indício contextual, como em D11 (seção 4.4), decorre da própria digitalização.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

5. CONSIDERAÇÕES FINAIS

Este relato examinou, em um caso real com conclusão técnica conhecida, até onde os metadados de arquivos PDF sustentam a detecção de indícios de adulteração em documentos nativos digitais apresentados em lote.

Quanto ao primeiro objetivo específico, a análise por arquivo atribuiu aos legítimos escore médio maior que aos não conformes, com ordenação próxima do acaso, por classificar a cadeia JasperReports/iText como software de edição. Ao perito, fica uma advertência prática: essa cadeia não constitui, por si só, indício de manipulação, pois é a assinatura de emissão dos sistemas institucionais que a utilizam, nem prova de autenticidade, pois quem falsifica também pode usá-la.

Quanto ao segundo, a correção da base de conhecimento elevou a AUC de 0,43 para 0,86, mas não alcançou os dois documentos compatíveis com fabricação integral. A separação veio do confronto entre o software produtor e uma referência externa ao arquivo: a origem institucional alegada separou todo o lote, e a cadeia predominante entre os pares alcançou seis dos sete não conformes, sem falso positivo, nos limiares de 1 a 15, e nenhum no limiar de 20. Neste lote, ambas equivalem a uma regra simples por arquivo (seção 4.6); os critérios relacionais foram redundantes com o perfil de emissor. Quanto ao terceiro, foram derivadas, sem teste no lote, fontes de falso positivo e vetores de evasão que delimitam o método ao documento original nativo digital.

A conclusão central é que os metadados de um arquivo, lidos isoladamente, indicam procedência, mas não bastam para afirmar autenticidade: o que torna o software produtor um indício é a origem que o documento alega ter. A contribuição do relato está menos na detecção, obtida em caso conhecido, do que na delimitação: o falso positivo sistemático de uma base de conhecimento que desconhece a emissão institucional, a insuficiência dos indícios de edição diante de documentos de exportação única e a dependência de uma referência externa ao arquivo. Como trabalho futuro, impõem-se a avaliação em conjuntos independentes, com documentos de emissão verificada; a construção de um conjunto público de comprovantes sintéticos, gerados com as cadeias de emissão reais e falsificados por pessoas que não conheçam os critérios, que permita avaliar o método com cegamento; a extração do tipo e do período a partir do conteúdo; a leitura das fontes tipográficas embarcadas e da camada XMP; e a identificação da ferramenta pelo conteúdo binário (Zia; Adedayo, 2025), que resiste à alteração direta dos metadados.

Os resultados não autorizam generalização nem comprovam a autenticidade ou a falsidade de nenhum documento. No uso pericial, os indícios da ferramenta servem à triagem;

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

para integrar um laudo avaliativo, precisariam ser expressos como razão de verossimilhança entre as hipóteses das partes, como orienta a diretriz europeia de relato avaliativo (European Network of Forensic Science Institutes, 2015).

A ferramenta produz indícios; a conclusão pericial é humana.

DECLARAÇÕES

Uso de inteligência artificial. Este trabalho utilizou a ferramenta Claude (Anthropic), modelo de linguagem generativo de uso comercial, entre setembro e novembro de 2026, para: levantamento bibliográfico preliminar, com os trabalhos identificados conferidos em registros bibliográficos oficiais e nas publicações originais; depuração, refatoração e testes do código da ferramenta; execução, sob supervisão dos autores, dos *scripts* de avaliação sobre os arquivos do caso e leitura do parecer, o que implicou o processamento desses arquivos e de seus metadados pelo serviço, nas condições descritas na seção 3.2; geração da figura; redação de versões preliminares das seções do texto, cada uma revisada e aprovada pelos autores; e revisão crítica do manuscrito. A concepção do trabalho, o exame que originou o caso, a definição dos critérios da ferramenta, a interpretação dos resultados e as conclusões são dos autores, que respondem integralmente pelo conteúdo. Nenhuma ferramenta de IA figura como autora.

Financiamento. Esta pesquisa não recebeu financiamento.

Conflito de interesses. Dois dos autores participaram do exame técnico que originou o caso relatado, um deles como perito signatário do parecer usado como padrão de referência (seção 3.7). Não há outros conflitos de interesses a declarar.

Ética. Pesquisa documental sobre material de exame técnico já concluído, dispensada de registro no sistema CEP/Conep com fundamento no art. 1º, parágrafo único, VII, da Resolução CNS nº 510/2016 (seção 3.2).

REFERÊNCIAS

ADHATARAO, S.; LAURADOUX, C. Exploitation and sanitization of hidden data in PDF files: do security agencies sanitize their PDF files? *In*: ACM WORKSHOP ON INFORMATION HIDING AND MULTIMEDIA SECURITY, 2021, [*S. l.*]. **Proceedings** [...]. New York: ACM, 2021. p. 35-44. DOI: 10.1145/3437880.3460405.

AITKEN, C. G. G.; TARONI, F. **Statistics and the evaluation of evidence for forensic scientists**. 2. ed. Chichester: Wiley, 2004. DOI: 10.1002/0470011238.

BRASIL. Conselho Nacional de Saúde. Resolução nº 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em Ciências Humanas e Sociais. **Diário Oficial da União**: seção 1, Brasília, DF, n. 98, p. 44-46, 24 maio 2016.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

BRASIL. Decreto-Lei nº 2.848, de 7 de dezembro de 1940. Código Penal. **Diário Oficial da União**: seção 1, Rio de Janeiro, 31 dez. 1940. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm. Acesso em: 26 out. 2026.

BRASIL. Decreto-Lei nº 3.689, de 3 de outubro de 1941. Código de Processo Penal. **Diário Oficial da União**: seção 1, Rio de Janeiro, 13 out. 1941. Disponível em: https://www.planalto.gov.br/ccivil_03/decreto-lei/del3689.htm. Acesso em: 26 out. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 24 out. 2026.

BRASIL. Ministério da Saúde. Comissão Nacional de Ética em Pesquisa. **Ofício Circular nº 17/2022/CONEP/SECNS/MS**: orientações acerca do artigo 1º da Resolução CNS nº 510, de 7 de abril de 2016. Brasília, DF, 5 jul. 2022. Disponível em: https://fcfrp.usp.br/wp-content/uploads/sites/528/2024/12/Oficio-Circular-17-2022.pdf. Acesso em: 3 nov. 2026.

BROWN, L. D.; CAI, T. T.; DASGUPTA, A. Interval estimation for a binomial proportion. **Statistical Science**, v. 16, n. 2, p. 101-133, 2001. DOI: 10.1214/ss/1009213286.

DEL PICCHIA FILHO, J.; DEL PICCHIA, C. M. R.; DEL PICCHIA, A. M. G. **Tratado de documentoscopia**: da falsidade documental. 3. ed. rev. ampl. atual. São Paulo: Pillares, 2016.

DROR, I. E. Cognitive and human factors in expert decision making: six fallacies and the eight sources of bias. **Analytical Chemistry**, v. 92, n. 12, p. 7998-8004, 2020. DOI: 10.1021/acs.analchem.0c00704.

EL JAKANI, Y.; RHAZI, M. N.; MEKNASSI, O.; CHBIKI, O.; MSAHLI, M.; GONNET, F. TrustPaper for forged PDF with issuer fingerprints. **arXiv**, 2026. Preprint. DOI: 10.48550/arXiv.2610.14132.

EUROPEAN NETWORK OF FORENSIC SCIENCE INSTITUTES. **ENFSI guideline for evaluative reporting in forensic science**: strengthening the evaluation of forensic results across Europe (STEOFRAE). Version 3.0. [*S. l.*]: ENFSI, 8 mar. 2015.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861-874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. **Radiology**, v. 143, n. 1, p. 29-36, 1982. DOI: 10.1148/radiology.143.1.7063747.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. **MIS Quarterly**, v. 28, n. 1, p. 75-105, 2004. DOI: 10.2307/25148625.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO/IEC 27037**: information technology: security techniques: guidelines for identification, collection, acquisition and preservation of digital evidence. Genebra: ISO, 2012.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 32000-1**: document management: portable document format: part 1: PDF 1.7. Genebra: ISO, 2008.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 32000-2**: document management: portable document format: part 2: PDF 2.0. Genebra: ISO, 2020.

JASPERSOFT. **Upgrade iText dependency**: issue 5. [*S. l.*]: GitHub, 28 mar. 2017. Disponível em: https://github.com/TIBCOSoftware/jasperreports/issues/5. Acesso em: 3 nov. 2026.

JASPERSOFT. **Configure the producer of the PDF document**: JSPR-I-4679. [*S. l.*]: Jaspersoft Ideas, 29 fev. 2024. Disponível em: https://jaspersoftideas.tibco.com/ideas/JSPR-I-4679. Acesso em: 24 out. 2026.

JASPERSOFT. **JasperReports 7.0.8**: release notes. [*S. l.*]: GitHub, 5 ago. 2026. Disponível em: https://github.com/Jaspersoft/jasperreports/releases/tag/7.0.8. Acesso em: 3 nov. 2026.

JASPERSOFT. **JasperReports 6.13.0**: descritor de dependências (POM). [*S. l.*]: Maven Central, [*s. d.*]a. Disponível em: https://repo1.maven.org/maven2/net/sf/jasperreports/jasperreports/6.13.0/jasperreports-6.13.0.pom. Acesso em: 3 nov. 2026.

JASPERSOFT. **jasperreports-pdf-lib7**. [*S. l.*]: GitHub, [*s. d.*]b. Disponível em: https://github.com/Jaspersoft/jasperreports-pdf-lib7. Acesso em: 22 out. 2026.

KASSIN, S. M.; DROR, I. E.; KUKUCKA, J. The forensic confirmation bias: problems, perspectives, and proposed solutions. **Journal of Applied Research in Memory and Cognition**, v. 2, n. 1, p. 42-52, 2013. DOI: 10.1016/j.jarmac.2013.01.001.

MAINKA, C.; MLADENOV, V.; ROHLMANN, S. Shadow attacks: hiding and replacing content in signed PDFs. *In*: NETWORK AND DISTRIBUTED SYSTEM SECURITY SYMPOSIUM, 28., 2021, [*S. l.*]. **Proceedings** [...]. Reston: Internet Society, 2021. DOI: 10.14722/ndss.2021.24117.

OLIVIER, M. S. On the classification of questioned PDF documents: attributing PDF documents to the tools that created them. **Forensic Science International: Digital Investigation**, v. 57, art. 302104, jun. 2026. DOI: 10.1016/j.fsidi.2026.302104.

PEFFERS, K.; TUUNANEN, T.; ROTHENBERGER, M. A.; CHATTERJEE, S. A design science research methodology for information systems research. **Journal of Management Information Systems**, v. 24, n. 3, p. 45-77, 2007. DOI: 10.2753/MIS0742-1222240302.

ROHLMANN, S.; MLADENOV, V.; MAINKA, C.; SCHWENK, J. Breaking the specification: PDF certification. *In*: IEEE SYMPOSIUM ON SECURITY AND PRIVACY, 42., 2021, San Francisco. **Proceedings** [...]. Los Alamitos: IEEE Computer Society, 2021. p. 1485-1501. DOI: 10.1109/SP40001.2021.00110.

SAKS, M. J.; KOEHLER, J. J. The coming paradigm shift in forensic identification science. **Science**, v. 309, n. 5736, p. 892-895, 2005. DOI: 10.1126/science.1111565.

TERRA INSIGHT. **Bank statement PDF metadata inspection**: what credit teams should check. [*S. l.*]: Terra Insight, 23 abr. 2026. Disponível em: https://www.terra-insight.com/insights/bank-statement-metadata-inspection/. Acesso em: 3 nov. 2026.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

WILSON, E. B. Probable inference, the law of succession, and statistical inference. **Journal of the American Statistical Association**, v. 22, n. 158, p. 209-212, 1927. DOI: 10.1080/01621459.1927.10502953.

WORSTER, A.; CARPENTER, C. Incorporation bias in studies of diagnostic tests: how to avoid being biased about bias. **CJEM**, v. 10, n. 2, p. 174-175, 2008. DOI: 10.1017/S1481803500009891.

ZIA, M. A. M.; ADEDAYO, O. M. Tool type identification for forensic digital document examination. **Forensic Science International: Digital Investigation**, v. 54, art. 301972, set. 2025. DOI: 10.1016/j.fsidi.2025.301972.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.