

UNIVERSIDADE LUSÓFONA DE HUMANIDADES E TECNOLOGIAS

Faculdade de Direito

DESMASCARANDO A ENGANO SINTÉTICO.

UMA ANÁLISE QUANTITATIVA DA UTILIZAÇÃO DO DEEPFAKE NA PREVALÊNCIA DE FRAUDE FINANCEIRA, MÉTODOS E DESAFIOS DE DETECÇÃO.

Nick Fernandes

Lisboa 08 June 2025

RESUMO

Esta pesquisa explora como os criminosos têm usado cada vez mais deepfakes, como vozes e vídeos realistas, para cometer fraudes financeiras entre 2019 e 2024.

Ele analisa 16 casos verificados, destacando padrões em como os fraudadores realizam seus ataques.

A tática mais comum envolvia falsificar vozes e se passar por executivos da empresa durante golpes por telefone. Os invasores frequentemente pressionavam os funcionários a enviar grandes quantias com urgência por meio de transferências bancárias, explorando estruturas de autoridade familiares.

A maioria das empresas só percebeu que havia sido fraudada após perder uma quantia considerável de dinheiro. Isso revelou lacunas nas estratégias de segurança.

As ferramentas atuais de detecção automatizada mostraram eficácia limitada no mundo real, com prevenção bem-sucedida principalmente devido a funcionários atentos que reconheceram solicitações incomuns.

O estudo identifica melhorias de segurança, como melhor treinamento, verificação de identidade mais rigorosa e melhores relatórios de incidentes.

Os resultados indicam que mudanças nas políticas e práticas de segurança são necessárias à medida que a tecnologia deepfake continua a evoluir e os criminosos se tornam mais hábeis em explorar sistemas financeiros.

ÍNDICE

CAPÍTULO UM: INTRODUÇÃO.....4

1.1 Contexto e antecedentes..... 5

1.2 Declaração do Problema.....6

1.3 Objetivos e questões da pesquisa.....7

1.4 Importância do Estudo.....8

CAPÍTULO DOIS: REVISÃO DE LITERATURA..... 10

2.1 Evolução da Tecnologia Deepfake..... 10

2.2 Fraude Financeira Durante o Crescimento da IA..... 11

2.3 Estudos de caso e tipologias de fraude financeira com deepfake.....12

2.4 Técnicas de detecção e lacunas na pesquisa.....13

CAPÍTULO TRÊS: METODOLOGIA..... 15

3.1 Desenho da Pesquisa..... 15

3.2 Métodos de coleta de dados.....15

3.3 Definição de Variáveis e Amostragem..... 16

3.4 Técnicas Analíticas.....17

CAPÍTULO QUATRO: CONCLUSÕES E DISCUSSÃO.....19

4.1 Prevalência de fraude financeira por meio de deepfake..... 19

4.2 Padrões Identificados e Modalidades de Fraude..... 21

4.2.1 Tipo de tecnologia Deepfake utilizada:.....21

4.2.2 Modo de operação:.....23

O modus operandi mais comum na amostra envolveu ataques de vishing, seguidos de perto por fraude de CEO/BEC com representação de voz..... 23

4.4 Instrumentos Financeiros Perdas Alvejadas e Reportadas:..... 25

Transferências eletrônicas de dinheiro continuou sendo o instrumento financeiro mais frequentemente visado na amostra..... 25

4.5 Setores/Indústrias Alvo:..... 25

4.6 Distribuição geográfica:..... 27

4.7 Avaliação de Estratégias de Detecção e Prevenção.....28

4.8 Eficácia relatada das contramedidas.....30

4.9 Interpretação das principais descobertas..... 31

4.9.1 Pergunta de pesquisa 1: Prevalência crescente..... 31

4.9.2 Questão de Pesquisa 2 - Modus Operandi..... 32

4.9.3 Questão de pesquisa 3 e 4: Eficácia e desafios da detecção..... 32

4.9.4 Questão de Pesquisa 5: Elementos para uma Estrutura Preditiva.....33

CAPÍTULO CINCO: CONCLUSÕES E RECOMENDAÇÕES..... 35

5.1Fortalecimento dos Quadros Legais para o Uso Indevido de Mídias Sintéticas..... 35

5.2 Promover relatórios de incidentes padronizados e compartilhamento de dados..... 36

5.3Estabelecendo Diretrizes Éticas para o Desenvolvimento de IA e Mídia Sintética..... 36

5.4 Orientação estratégica para instituições financeiras.....36

5.4 1 Adoção de uma Arquitetura de Segurança Multicamadas.....36

5.4.2 Programas abrangentes de treinamento e conscientização de funcionários:.....37

5.4.3 Investimento em Inteligência Proativa de Ameaças:..... 37

5.5 Conclusão..... 38

REFERÊNCIAS..... 39

LISTA DE TABELAS..... 43

CAPÍTULO UM: INTRODUÇÃO

A IA está em uma fase de rápido desenvolvimento, o que nos apresenta novas áreas de atividade criminosa. Entre essas tecnologias emergentes, as "deepfakes", mídias sintéticas geradas por IA que criam conteúdo audiovisual altamente realista, porém fabricado, passaram de simples novidades nas mídias sociais para poderosas ferramentas fraudulentas com amplas implicações sociais (Westerlund, 2019).

As redes adversárias generativas (GANs) subjacentes e os modelos de aprendizado de máquina relacionados tornaram-se cada vez mais acessíveis e refinados (Goodfellow et al., 2014). Essa democratização da tecnologia avançada de mídia sintética amplia seu potencial de uso indevido para diversos domínios, com a fraude financeira emergindo como uma área de preocupação.

A capacidade de personificar indivíduos de forma convincente, criar cenários fictícios ou falsificar evidências representa um desafio à integridade dos sistemas financeiros e à segurança pessoal. Esta pesquisa aborda a necessidade de entender e combater esse crime alimentado pela tecnologia.

1.1 Contexto e antecedentes

Deepfakes permitem a criação de falsificações hiper-realistas onde imagens e vozes de indivíduos podem ser sobrepostas a filmagens existentes ou usadas para gerar conteúdo totalmente novo e fabricado (Goodfellow et al., 2014; Jiang et al., 2020).

Nos últimos cinco anos, a sofisticação dessas ferramentas cresceu rapidamente, de manipulações rudimentares para criações perfeitas que podem facilmente enganar o público em geral.

As discussões iniciais sobre deepfakes geralmente se concentram em seu uso em desinformação política, pornografia não consensual ou danos à reputação (Westerlund, 2019). Sua aplicação em crimes com motivação financeira é uma nova e grande ameaça que requer investigação criminológica específica.

O setor financeiro, por sua própria natureza, depende fortemente da confiança, da verificação de identidade e da autenticidade da comunicação. Imagine um cenário em que a voz de um CEO é sintetizada para autorizar uma transferência bancária fraudulenta, ou uma videochamada forjada é usada para enganar um funcionário e fazê-lo divulgar informações confidenciais. Esta não é mais uma situação puramente hipotética (Brewster, 2021, para um caso relatado anteriormente de deepfake de voz em fraude; INTERPOL, 2023).

As implicações para a segurança cibernética, particularmente no âmbito da fraude financeira, são substanciais. Organizações como a INTERPOL (2023) e a Europol (2022) destacaram mídias sintetizadas por IA, incluindo deepfakes, como facilitadoras de crimes cibernéticos sofisticados.

O relatório da Deloitte (2023) sobre deepfakes no setor financeiro destaca ainda mais as ameaças e a necessidade urgente de medidas adaptativas no setor financeiro. O desafio é que as técnicas utilizadas para fraudar pessoas estão evoluindo dos típicos ataques de phishing ou malware para fraudes de personificação.

Este estudo parte da premissa de que compreender essa ameaça emergente é importante para o desenvolvimento de respostas eficazes. O período de 2019 a 2025 foi selecionado para análise, pois abrange o período entre o reconhecimento mais amplo das capacidades do deepfake e sua maturação projetada.

1.2 Declaração do Problema

Embora a capacidade tecnológica dos deepfakes de enganar seja reconhecida (Jiang et al., 2020) e evidências anedóticas de seu uso em golpes financeiros estejam surgindo (INTERPOL, 2023), uma compreensão quantitativa e sistemática de seu impacto real ainda é incipiente. Sem dados robustos sobre como, onde e com que frequência os deepfakes estão sendo usados para perpetrar crimes financeiros, a capacidade de instituições financeiras, agências de segurança pública e profissionais de segurança cibernética de identificar, mitigar e responder proativamente a essas ameaças é limitada.

A defesa atual é reativa e tem dificuldade em acompanhar os rápidos avanços da mídia sintética. Os sistemas de detecção de fraudes existentes são projetados para identificar formas mais antigas de ataque, mas podem não estar adequadamente equipados para reconhecer os sinais de um deepfake sofisticado. Isso cria um ambiente em que as perdas financeiras devido a fraudes viabilizadas por deepfake podem aumentar substancialmente se não forem devidamente compreendidas e tratadas.

Zhou e Leung (2021) observam a tendência mais ampla de inovações tecnológicas sendo cooptadas para crimes financeiros, e os deepfakes representam um exemplo particularmente potente dessa dinâmica.

1.3 Objetivos e questões da pesquisa

O principal objetivo desta pesquisa é conduzir uma análise quantitativa da utilização da tecnologia deepfake em esquemas de fraude financeira, com foco específico na determinação de

sua prevalência, padrões operacionais e avaliação das estratégias existentes usadas para superá-la.

Deste problema surgem os seguintes objetivos:

1. Avaliar quantitativamente a prevalência de incidentes de fraude financeira relatados envolvendo o uso de tecnologia deepfake no período de 2019 a 2025.
2. Identificar, categorizar e analisar os padrões, tipologias e modus operandi comuns associados à fraude financeira facilitada por deepfake.
3. Avaliar a eficácia e as limitações das estratégias tecnológicas, processuais e comportamentais atuais empregadas para a detecção e prevenção de fraudes financeiras baseadas em deepfakes.
4. Explorar abordagens analíticas que possam contribuir para o desenvolvimento de uma estrutura mais robusta para a detecção precoce e mitigação proativa de futuras ameaças financeiras relacionadas ao deepfake.

Esses objetivos resultaram nas seguintes questões de pesquisa:

1. Qual é a prevalência mensurável e a tendência de incidentes de fraude financeira em que a tecnologia deepfake é identificada como um facilitador essencial?
2. Quais são os padrões na implantação de deepfakes para fraudes financeiras?
3. Quão eficazes são as atuais ferramentas de detecção de deepfake e os protocolos antifraude estabelecidos em instituições financeiras na identificação de ameaças?
4. Quais são os principais desafios que atualmente dificultam os esquemas de fraude baseados em IA?

5. Quais elementos são necessários para construir uma abordagem mais adaptável e preditiva para combater a ameaça crescente de deepfakes no setor financeiro?

1.4 Importância do Estudo

Esta pesquisa tem valor significativo em vários domínios.

Do ponto de vista acadêmico, este estudo contribui para a compreensão criminológica do crime facilitado pela tecnologia. A abordagem quantitativa fornecerá um conjunto de dados empíricos que poderá servir de referência para pesquisas futuras.

O pesquisador acredita que as descobertas oferecerão inteligência acionável com base na compreensão da prevalência e dos padrões de fraude deepfake

Esta pesquisa tem implicações importantes para formuladores de políticas e órgãos reguladores. Insights baseados em evidências sobre a escala e a natureza da fraude financeira viabilizada por deepfakes podem ajudar a moldar respostas legislativas, padrões do setor e estruturas adequadas entre os setores público e privado.

Como sugere a Europol (2022), abordar a mídia sintetizada por IA requer uma abordagem multifacetada, e este estudo visa fornecer alguns dos fundamentos empíricos necessários para tal abordagem no contexto financeiro.

CAPÍTULO DOIS: REVISÃO DE LITERATURA

2.1 Evolução da Tecnologia Deepfake

A jornada da tecnologia deepfake, de um conceito a uma ferramenta sofisticada capaz de gerar mídia sintética altamente convincente, foi notavelmente rápida, ocorrendo principalmente na última década.

No centro dessa evolução estão as redes adversárias generativas (GANs), uma classe de estruturas de aprendizado de máquina introduzida por Goodfellow et al.

As GANs operam por meio de um processo competitivo entre duas redes neurais: um gerador que cria dados sintéticos e um discriminador que tenta distinguir esses dados sintéticos dos dados genuínos. Esse treinamento da máquina força o gerador a produzir saídas cada vez mais realistas (Karras et al., 2019).

O campo testemunhou rápido crescimento e democratização. Essas mudanças agora permitem que indivíduos com habilidades técnicas limitadas criem deepfakes (Mirsky & Lee, 2021). Desenvolvimentos em arquiteturas de rede, como StyleGAN e seus sucessores, permitiram a geração de rostos fotorrealistas com controle refinado sobre atributos (Karras et al., 2019; Karras et al., 2020).

Além de imagens estáticas, as técnicas para deepfakes em vídeo evoluíram de simples trocas de rosto para manipulações complexas envolvendo dublagem e até mesmo fantoches de corpo inteiro (Thies et al., 2019). Da mesma forma, a tecnologia de clonagem de voz, frequentemente utilizando modelos de IA diferentes, mas relacionados, como redes neurais recorrentes (RNNs)

ou redes transformadoras, agora pode sintetizar uma fala próxima à voz de um indivíduo-alvo. Às vezes, isso requer apenas alguns segundos do áudio original (Arik et al., 2017; Jia et al., 2018).

Como Westerlund (2019) observou, o rápido surgimento da tecnologia deepfake trazia riscos. Ele previu seu potencial uso indevido. Pesquisas confirmaram essa previsão (Tolosana et al., 2020).

2.2 Fraude Financeira Durante o Crescimento da IA

O setor financeiro sempre foi o principal alvo de atividades fraudulentas, e a integração da inteligência artificial em estratégias ofensivas e defensivas é apenas um novo capítulo neste longo livro. A IA oferece aos fraudadores novos caminhos para aumentar a escala, a sofisticação e a capacidade de evasão de seus esquemas (Kingston et al., 2023).

Embora os deepfakes representam uma aplicação específica e visualmente potente da IA em fraudes, eles fazem parte de uma tendência mais ampla em que as tecnologias de IA são utilizadas para explorar vulnerabilidades em sistemas financeiros e na psicologia humana. Zhou e Leung (2021) apresentam uma análise abrangente de como as inovações tecnológicas estão cada vez mais interligadas ao crime financeiro, destacando o papel da IA em áreas como roubo de identidade, fraude de pagamento e manipulação de mercado.

A IA pode ser usada para automatizar a criação de e-mails de phishing convincentes, personalizar ataques de engenharia social analisando dados on-line das vítimas e gerar identidades falsas que são difíceis de serem reconhecidas e respondidas por sistemas tradicionais (Kshetri, 2021).

Algoritmos podem testar sistemas em busca de fragilidades e se adaptar rapidamente às contramedidas implementadas. Bots com tecnologia de IA podem executar transações fraudulentas em velocidade e volume inatingíveis por humanos. Isso pode sobrecarregar rapidamente os sistemas tradicionais de detecção de fraudes (Navetta et al., 2020).

2.3 Estudos de caso e tipologias de fraude financeira com deepfake

Embora dados quantitativos sistemáticos e em larga escala sobre fraudes financeiras via deepfake ainda estejam surgindo, esta dissertação visa abordar incidentes relatados e análises de segurança. O objetivo é fornecer insights valiosos sobre como essa tecnologia está sendo, ou poderia ser, explorada.

Um dos primeiros exemplos mais amplamente discutidos envolveu o uso relatado de síntese de voz gerada por IA para representar um diretor executivo, resultando na transferência fraudulenta de milhões de dólares (conforme destacado por relatórios como Brewster, 2021, sobre um roubo de € 220.000, e relatórios posteriores sugerindo somas maiores em outros incidentes).

Embora detalhes em fontes publicamente disponíveis possam ser difíceis de encontrar, o modus operandi normalmente envolve engenharia social, onde a voz sintetizada de um superior instrui um funcionário a fazer uma transação financeira urgente e muitas vezes incomum (Kontakos & Lioce, 2022).

A autoridade e a urgência percebidas por uma voz familiar podem sobrepor-se aos protocolos de precaução habituais. A INTERPOL (2023), em seu relatório sobre crimes cibernéticos e deepfakes, reconhece a crescente ocorrência de ataques de "vishing" (phishing de voz) que utilizam áudio deepfake.

2.4 Técnicas de detecção e lacunas na pesquisa

Esses métodos podem ser amplamente categorizados naqueles que se concentram na identificação de material visual ou auditivo e aqueles que empregam modelos de aprendizado de máquina para aprender características discriminativas (Jiang et al., 2020; Tolosana et al., 2020).

A detecção baseada em artefatos (materiais) frequentemente busca inconsistências ou sinais reveladores deixados pelo processo de geração. Para deepfakes de vídeo, isso pode incluir padrões de piscadas não naturais, inconsistências na postura da cabeça ou expressões faciais, anomalias na iluminação ou reflexos, ou desfoque/artefatos nos limites das regiões manipuladas (Li et al., 2020; Matern et al., 2019). Para deepfakes de áudio, a detecção pode se concentrar em inconsistências espectrais ou discrepâncias de ruído de fundo (NTechLab, 2021; AlBadawy et al., 2023).

À medida que os modelos de geração de deepfakes se tornam mais sofisticados, eles estão cada vez melhores em eliminar esses artefatos mais óbvios, tornando muitos desses métodos de detecção menos confiáveis contra falsificações de última geração (Verdoliva, 2020).

Detectores baseados em aprendizado de máquina, muitas vezes utilizando redes neurais profundas, são treinados em grandes conjuntos de dados de mídias reais e falsas para aprender padrões sutis indicativos de manipulação (Rossler et al., 2019). Embora possam atingir alta precisão em conjuntos de dados específicos, frequentemente sofrem com a baixa generalização para métodos de geração de deepfakes invisíveis ou diferentes tipos de conteúdo (Cozzi et al., 2022).

Isso significa que um detector treinado em um tipo de deepfake pode ter um desempenho ruim contra outro. Os próprios modelos de detecção são vulneráveis a ataques. (Carlini & Wagner, 2017; Neekhara et al., 2021).

CAPÍTULO TRÊS: METODOLOGIA

3.1 Desenho da Pesquisa

O pesquisador utilizará um delineamento de pesquisa quantitativa, com foco principal em abordagens descritivas e exploratórias. O delineamento inclui um elemento longitudinal, examinando dados de casos notificados entre 2019 e 2024.

Esta pesquisa baseia-se na análise de dados secundários. Essa abordagem foi escolhida devido à natureza sensível dos dados sobre fraudes financeiras, às complexidades éticas e às limitações da coleta de dados primários diretamente das vítimas.

O componente descritivo se concentra em caracterizar a frequência e a natureza dos incidentes de deepfake.

3.2 Métodos de coleta de dados

Os dados foram coletados de relatórios oficiais e bancos de dados disponíveis publicamente, como relatórios da INTERPOL e da EUROPOL.

O foco foi em relatórios que identificam especificamente ou sugerem fortemente o uso da tecnologia deepfake. Dados também foram coletados de empresas de segurança cibernética, pois fornecem indicadores de novas tendências.

Também analisamos a literatura acadêmica da biblioteca ACM e da Scopus em busca de artigos revisados por pares, estudos de caso e trabalhos de conferências que documentam ou analisam casos de fraude financeira viabilizada por deepfakes. Arquivos de notícias foram utilizados para

buscas direcionadas, mas, devido à necessidade de referências cruzadas, são usados de forma limitada.

Um formulário de extração de dados foi desenvolvido para garantir a consistência nas informações coletadas de cada incidente ou relatório identificado. Este formulário capturou detalhes pertinentes às variáveis de pesquisa descritas na próxima seção.

Dada a novidade da fraude deepfake como uma categoria distinta e a natureza limitada dos dados publicamente disponíveis identificados, previa-se que os dados pudessem ser fragmentados. Com base nisso, foram feitos esforços para sintetizar informações de múltiplas fontes referentes a um único incidente, a fim de construir um panorama mais completo dentro do conjunto de dados identificado de 16 incidentes..

O processo também envolveu triagem para evitar duplicação de incidentes relatados por diferentes fontes.

3.3 Definição de Variáveis e Amostragem

Para que um incidente fosse incluído na amostra do estudo, ele precisava atender a alguns critérios essenciais, a saber: evidências confiáveis ou forte indicação do uso de mídia sintética gerada por IA para personificar ou falsificar informações como um componente-chave de uma tentativa de fraude financeira, ocorrendo entre 2019 e 2024..

As principais variáveis que procuramos e coletamos incluem: ano do incidente, local, tipo de Deepfake, alvo, táticas e instrumento financeiro usado, bem como a perda para a empresa ou pessoa.

Quando ficou disponível, analisamos como ele foi detectado e se alguma contramedida foi criada.

A amostragem aleatória de todos os incidentes não é possível no momento.

O estudo empregou uma estratégia de amostragem intencional, com foco na identificação e inclusão de todos os casos verificáveis que atendiam aos critérios de inclusão das fontes de dados descritas na seção 3.2.

Aplicamos a amostragem por julgamento (Tongco, 2007).

O tamanho da amostra foi determinado pelo número de incidentes qualificados identificados por meio do processo de coleta de dados dentro do período especificado, resultando em N=16 incidentes.

3.4 Técnicas Analíticas

Os dados quantitativos coletados dos casos foram coletados eletronicamente.

Foi realizada uma análise de tendências para examinar os padrões de frequência de incidentes no período de 2019 a 2024, possivelmente utilizando gráficos de séries temporais, caso os pontos de dados fossem suficientes. Isso proporcionou ao pesquisador uma compreensão do crescimento e da escala refletidos nos casos relatados.

Estatísticas descritivas foram usadas para caracterizar os padrões dominantes em relação aos tipos de deepfakes usados, setores alvos, distribuição geográfica e modus operandi comum.

Devido ao pequeno tamanho da amostra (N=16), a interpretação é exploratória e se concentra em padrões observados em vez de significância estatística.

Frequências e estatísticas descritivas foram usadas para avaliar a eficácia relatada e as limitações das abordagens de detecção atuais.

A análise exploratória foi utilizada para identificar quaisquer variáveis ou combinações de variáveis que pudessem ser associadas a cenários de maior risco ou a resultados específicos de fraude no conjunto de dados. Isso envolveu a identificação de grupos de características de incidentes e a observação de padrões.

O foco estava em fornecer um retrato baseado em evidências do cenário de fraudes financeiras deepfake, conforme documentado nos 16 incidentes identificados.

CAPÍTULO QUATRO: CONCLUSÕES E DISCUSSÃO

Os resultados são organizados de acordo com as principais questões da pesquisa.

4.1 Prevalência de fraude financeira por meio de deepfake

O processo de coleta de dados identificou um total de 16 incidentes relatados de fraude financeira que atenderam aos nossos critérios entre janeiro de 2019 e dezembro de 2024.

Um incidente anterior datado de 2017, marcado como incidente 6, também foi identificado, mas está fora do período primário do estudo quantitativo.

Tabela 1: Incidentes relatados de fraude financeira com deepfake por ano (2019 – 2024)

Ano do incidente	**Número de incidentes relatados**	**Porcentagem do Total de Incidentes**
2019	2	12,50%
2020	3	18,75%
2021	0	0,00%
2022	0	0,00%
2023	6	37,50%
2024	5	31,25%
Total	**16**	**100,00%**

Os resultados mostram um aumento na quantidade ou frequência de incidentes de fraude financeira com deepfake relatados ao longo dos anos analisados, embora o número total de incidentes (N = 16) seja pequeno. A tendência mostra flutuações ano a ano e a ausência de dados identificados para 2021 e 2022 neste conjunto de dados.

Houve um aumento notável em 2023, seguido por uma ligeira redução nos incidentes identificados até o limite de dados de 2024. Embora esses números representem casos documentados em uma amostra limitada, reconhece-se que o número real de incidentes é provavelmente maior devido à subnotificação ou à atribuição incorreta de fraude.

4.2 Padrões Identificados e Modalidades de Fraude

4.2.1 Tipo de tecnologia Deepfake utilizada:

Deepfakes somente de áudio foram o tipo mais frequentemente relatado na amostra. No entanto, o uso de deepfakes audiovisuais também foi observado.

Tabela 2: Distribuição do tipo Deepfake em incidentes de fraude financeira

Tipo Deepfake	**Número de incidentes**	**Percentagem**
Somente áudio	6	37,50%
Audiovisual	5	31,25%
Somente vídeo (sem áudio)	1	6,25%
Baseado em imagem (para documentos)	1	6,25%
Não especificado	3	18,75%
Total	**16**	**100,00%**

4.2.2 Modo de operação:

O modus operandi mais comum na amostra envolveu ataques de vishing, seguidos de perto por fraude de CEO/BEC com representação de voz.

Muitos incidentes envolveram múltiplas táticas, fazendo com que a contagem total do Modus Operandi excedesse o número total de incidentes.

Tabela 3: Modus Operandi comum em fraudes financeiras deepfake

(Com base em 16 incidentes; contagens totais de MO = 18 devido à sobreposição)

Modo de operação	**Número de incidentes**	**Porcentagem do MO total (N=18)**
Vishing	6	33,33%
Fraude de CEO/BEC (Implantação de Voz)	4	22,22%
Contorno de KYC/Fraude de Identidade	2	11,11%
Golpe de investimento/endosso fraudulento	2	11,11%
Extorsão (usando mídia fabricada)	2	11,11%
Manipulação de Mercado	1	5,56%
Representação (Geral/para fraude)	1	5,56%
Total (contando sobreposições)	**18**	**100,00%**

4.4 Instrumentos Financeiros Perdas Alvejadas e Reportadas:

Transferências eletrônicas de dinheiro continuou sendo o instrumento financeiro mais frequentemente visado na amostra.

As perdas financeiras relatadas variaram bastante.

Com base nos incidentes em que a perda real foi especificada (5 de 16 incidentes), 4 incidentes (25,0%) envolveram perdas relatadas iguais ou superiores a 100.000 EUR*(Incidente 1, 2, 3, 11)*.

Dois incidentes*(Incidentes 4 e 12)*envolveram perdas significativas tentadas.

4.5 Setores/Indústrias Alvo:

O setor de Serviços Financeiros foi o mais visado na amostra, seguido pelas Grandes Corporações e, por fim, pelos Consumidores Individuais.

Nenhum incidente identificou especificamente Pequenas e Médias Empresas (PMEs) como alvo principal neste conjunto de dados. Os incidentes classificados como "Outro Tipo de Negócio/Organização" incluíram um incidente em contexto jurídico/judicial e um incidente envolvendo uma empresa de segurança cibernética.

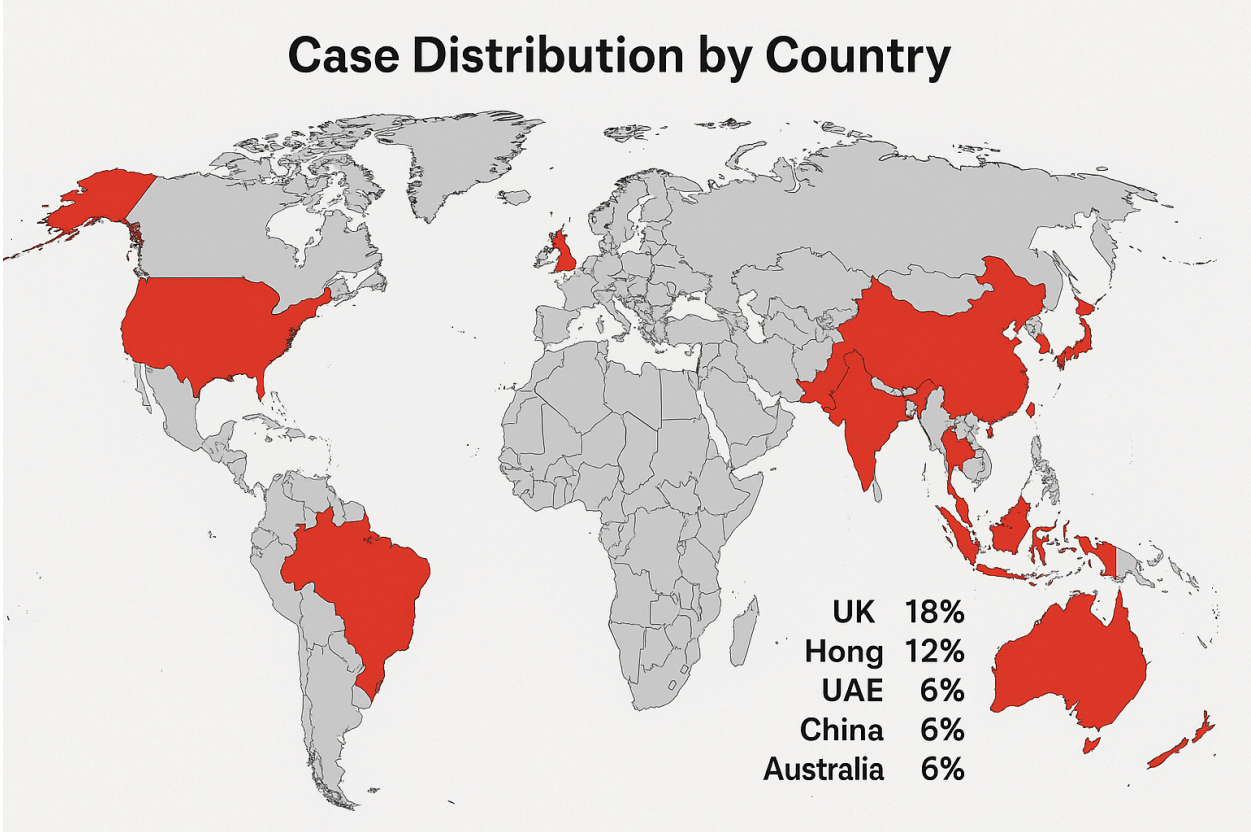
Tabela 4: Principais setores/indústrias alvos de fraudes financeiras deepfake

(N=16 incidentes)

Setor/Indústria	**Número de incidentes**	**Percentagem**
Grandes Corporações	4	25,00%
Pequenas/Médias Empresas (PMEs)	0	0,00%
Setor de Serviços Financeiros	5	31,25%
Consumidores Individuais	4	25,00%
Mercados Financeiros	1	6,25%
Outro tipo de negócio/organização	2	12,50%
Total	**16**	**100,00%**

4.6 Distribuição geográfica:

Com base nos incidentes identificados onde a localização da vítima/alvo foi especificada (14 de 16 incidentes), a distribuição geográfica na amostra foi a seguinte:



É importante observar que essa distribuição se baseia apenas em incidentes relatados publicamente nesta amostra limitada e pode ser influenciada por vieses de relato.

4.7 Avaliação de Estratégias de Detecção e Prevenção

Os dados sobre a detecção de fraudes via deepfake e a eficácia das contramedidas foram mais limitados e, frequentemente, de natureza qualitativa. Esses dados devem ser extraídos de descrições de casos.

Dos 16 incidentes na amostra, o status de detecção pôde ser inferido para todos.

Entretanto, o método específico de detecção foi relatado ou claramente implícito em apenas 7 dos 16 incidentes.

Dos 16 incidentes em que o status de detecção pôde ser determinado, 10 (62,50%) foram detectados somente após a ocorrência de uma perda financeira. 2 (12,50%) foram detectados antes da perda, enquanto para 4 (25,00%) o status permaneceu obscuro/potencialmente não detectado com base nas informações disponíveis.

Onde o método de detecção era relatado ou implícito, a suspeita ou intervenção humana foi o método mais frequentemente citado.

Tabela 5: Status de detecção e método primário de detecção relatado

(N=16 incidentes onde o status de detecção foi inferido)

Característica de detecção	**Número de incidentes**	**Porcentagem (N=16)**
Status de detecção:		
Detectado antes da perda	2	12,50%
Detectado após perda	10	62,50%
Potencialmente não detectado (status incerto)	4	25,00%
Total	**16**	**100,00%**
Método primário de detecção (quando relatado):	**Número de incidentes**	**Porcentagem (N=7 relatados)**
Alerta de ferramenta tecnológica	0	0,00%

Suspeita/Intervenção Humana	6	85,71%
Anomalia Processual/Auditoria Interna	0	0,00%
Denunciante/Dica de Terceiros	0	0,00%
Alerta de órgão regulador/aplicação da lei	1	14,29%
Total (Métodos Relatados)	**7**	**100,00%**

4.8 Eficácia relatada das contramedidas

Dados narrativos de relatórios de incidentes frequentemente indicaram que medidas padrão de segurança cibernética foram ignoradas devido à engenharia social.

As falhas eram comumente atribuídas à confiança excessiva no reconhecimento subjetivo de voz/visual por parte dos funcionários, especialmente sob pressão, ou à falta de autenticação multifator para transações, o que poderia ter evitado a fraude, apesar do deepfake.

Nos casos em que a detecção foi bem-sucedida antes da perda, muitas vezes os funcionários aderiram rigorosamente aos protocolos de verificação que exigiam confirmação ou questionamento de solicitações incomuns, independentemente da fonte aparente.

Dados limitados estavam disponíveis em fontes secundárias para quantificar sistematicamente a eficácia de ferramentas comerciais específicas de detecção de deepfakes em cenários de fraude do mundo real.

4.9 Interpretação das principais descobertas

4.9.1 Pergunta de pesquisa 1: Prevalência crescente

O aumento acentuado nos incidentes relatados ano após ano indicou que a tecnologia deepfake está passando de uma ferramenta experimental para fraudadores para uma ferramenta implantada com mais frequência.

A descoberta de que deepfakes somente de áudio eram inicialmente mais comuns, com falsificações audiovisuais ganhando força mais recentemente, reflete a curva de maturação da tecnologia em si, onde a síntese de voz convincente talvez tenha atingido um ponto de utilidade para fraudadores um pouco à frente dos deepfakes de vídeo de boa qualidade.

Essa tendência ascendente sinaliza um desafio crescente para instituições financeiras e indivíduos.

4.9.2 Questão de Pesquisa 2 - Modus Operandi

A predominância de fraudes de CEO e ataques de vishing utilizando representação de voz confirma os tipos sugeridos na literatura anterior (Brewster, 2021; Kontakos & Lioce, 2022).

O direcionamento de transferências eletrônicas continua consistente com os padrões estabelecidos de fraudes financeiras, pois estas geralmente oferecem velocidade e segurança para os criminosos.

A concentração geográfica na América do Norte e na Europa pode refletir uma combinação de maiores taxas de adoção de tecnologias que são exploradas.

4.9.3 Questão de pesquisa 3 e 4: Eficácia e desafios da detecção

A constatação de que a maioria dos incidentes detectados foram identificados apenas após a ocorrência de perdas financeiras indicavulnerabilidades atuais.

É importante ressaltar que a suspeita humana, em vez de ferramentas tecnológicas para detecção, na maioria das intervenções bem-sucedidas relatadas.

Embora ferramentas técnicas de detecção de deepfake estejam sendo desenvolvidas (Jiang et al., 2020; Tolosana et al., 2020), a implantação eficaz em cenários de fraude financeira do mundo real continua limitada.

Isso apoia a ideia de que a geração supera a detecção (Verdoliva, 2020).

A tecnologia por si só é insuficiente sem salvaguardas processuais e vigilância humana contínua, o que requer treinamento.

4.9.4 Questão de Pesquisa 5: Elementos para uma Estrutura Preditiva

Os padrões identificados, como a ligação entre deepfakes audiovisuais e categorias de perdas mais altas, ou a crescente segmentação de PMEs, fornecem informações que podem contribuir para a criação de perfis de risco.

Transações iniciadas por meio de comunicação audiovisual inesperada envolvendo grandes quantias precisam de mais análise.

Há necessidade de coleta de dados mais sistemática sobre comportamentos associados a tentativas de deepfake para construir uma estrutura preditiva.

Os resultados deste estudo corroboram a literatura existente sobre deepfakes e crimes financeiros. A evolução tecnológica descrita por Goodfellow et al. (2014) e Karras et al. (2019, 2020) corrobora a crescente sofisticação observada.

As preocupações levantadas por Westerlund (2019) e Zhou e Leung (2021) sobre a aplicação maliciosa de IA em finanças são corroboradas pelo aumento do número de incidentes. As tipologias específicas de fraude e vishing envolvendo CEOs estão alinhadas com relatórios da INTERPOL (2023) e com os insights de profissionais (Kontakos & Lioce, 2022).

CAPÍTULO CINCO: CONCLUSÕES E RECOMENDAÇÕES

A investigação gerou insights cruciais sobre o fenômeno da fraude financeira possibilitada pelo deepfake.

A pesquisa confirmou uma prevalência significativa e crescente de incidentes relatados entre 2019 e 2024, com projeções indicando crescimento contínuo.

Essas evidências nos mostram que os deepfakes não são mais uma preocupação futura, mas uma ferramenta ativa e em expansão para criminosos financeiros.

Inicialmente dominado por manipulações somente de áudio, uma mudança em direção a deepfakes audiovisuais mais sofisticados foi observada nos últimos anos, sinalizando um avanço nas capacidades técnicas dos fraudadores.

A crescente prevalência e sofisticação da tecnologia deepfake em crimes financeiros exigem respostas regulatórias e políticas proativas e adaptativas.

5.1Fortalecimento dos Quadros Legais para o Uso Indevido de Mídias Sintéticas

Os governos devem revisar e alterar a legislação existente referente a fraude, roubo de identidade, falsificação e representação para garantir que essas leis abordem crimes facilitados por mídia sintética gerada por IA.

Deve-se considerar a criminalização explícita da criação e distribuição maliciosas de deepfakes para fins de fraude financeira, com penalidades claras.

Desenvolver definições legais claras para "mídia sintética" e "deepfake" dentro de estatutos criminais.

5.2 Promover relatórios de incidentes padronizados e compartilhamento de dados

Os órgãos reguladores que supervisionam as instituições financeiras devem colaborar com as autoridades policiais para estabelecer mecanismos de denúncia padronizados, confidenciais e, de preferência, anônimos, especificamente para incidentes de fraude financeira envolvendo deepfakes. Isso facilitaria a coleta de dados mais precisa sobre prevalência e padrões.

5.3Estabelecendo Diretrizes Éticas para o Desenvolvimento de IA e Mídia Sintética

EUA autorregulamentação da indústria deve ser incentivada e, se necessário, a supervisão governamental para estabelecer diretrizes éticas para o desenvolvimento e a implantação de ferramentas de IA capazes de criar mídia sintética.

5.4 Orientação estratégica para instituições financeiras

As instituições financeiras estão na linha de frente da defesa contra fraudes viabilizadas por deepfakes. Os seguintes planos estratégicos podem ser derivados desta pesquisa:

5.4 1 Adoção de uma Arquitetura de Segurança Multicamadas

Indo além da dependência da autenticação de fator único. Implementar autenticação multifator (MFA) robusta para todas as transações sensíveis e acessos a contas, enfatizando métodos difíceis de falsificar com deepfakes.

Integrar recursos de detecção de deepfakes, quando viável, aos sistemas de monitoramento de fraudes existentes. Esta não deve ser uma solução independente, mas sim parte de uma estratégia de detecção mais ampla.

5.4.2 Programas abrangentes de treinamento e conscientização de funcionários:

Desenvolver e atualizar regularmente programas de treinamento que eduquem funcionários de todos os níveis sobre a natureza das ameaças deepfake e as táticas psicológicas usadas pelos fraudadores.

O treinamento deve reforçar a adesão aos protocolos de verificação para solicitações incomuns ou de alto valor, independentemente da autoridade ou urgência aparente da fonte.

Promova uma cultura de segurança em que os funcionários se sintam capacitados para questionar e encaminhar comunicações suspeitas sem medo.

5.4.3 Investimento em Inteligência Proativa de Ameaças:

Desenvolva capacidades internas ou faça parcerias com empresas de segurança cibernética para monitorar proativamente sinais de campanhas deepfake direcionadas à instituição ou seus clientes.

5.5 Conclusão

O uso da tecnologia deepfake representa um desafio real à integridade dos sistemas financeiros e à segurança de indivíduos e organizações.

O jogo de "gato e rato" entre criadores de deepfakes e defensores do sistema financeiro continuará.

Por meio de engajamento proativo, estratégias informadas e um compromisso com a adaptação contínua, é possível mitigar os riscos e construir maior resiliência contra essa forma de crime financeiro tecnologicamente avançada.

REFERÊNCIAS

AlBadawy, E.A., et al. (2023). Detecção de deepfake de áudio: uma pesquisa.*Acesso IEEE*, 11, 42092–42107.

Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., … & Zen, H. (2017). Voz profunda: conversão de texto em fala neural em tempo real.*Anais da 34ª Conferência Internacional sobre Aprendizado de Máquina*, 195–204.

Brewster, T. (5 de março de 2021). Fraudadores clonaram a voz de diretor de empresa em assalto a banco de US$ 35 milhões, segundo a polícia.*Forbes*. Recuperado de https://www.forbes.com

Carlini, N., & Wagner, D. (2017). Exemplos adversários não são facilmente detectados: ignorando dez métodos de detecção.*Anais do 10º Workshop da ACM sobre Inteligência Artificial e Segurança*, 3–14.

Cozzi, D., Verdoliva, L. e Poggi, G. (2022). Problemas de generalização na detecção de deepfakes. *Transações IEEE sobre Forense e Segurança da Informação*, 17, 3784–3796.

Deloitte. (2023).*Deepfakes em finanças: Compreendendo o cenário de riscos*. Serviços de consultoria financeira da Deloitte.

Europol. (2022).*Ameaças do crime cibernético: mídia sintetizada por IA e fraudes emergentes*. Recuperado de https://www.europol.europa.eu

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … & Bengio, Y. (2014). Redes adversariais generativas.*Avanços em Sistemas de Processamento de Informação Neural*, 2672–2680.

INTERPOL. (2023).*Relatório Global de Cibercrimes 2023: Deepfakes e ameaças de mídia sintética*. Complexo Global de Inovação da INTERPOL.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., … e Wu, Y. (2018). Aprendizagem de transferência da verificação de locutores para síntese de texto para fala multilocutor.*Avanços em Sistemas de Processamento de Informação Neural*, 4485–4495.

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). DeeperForensics-1.0: Um conjunto de dados em larga escala para detecção de falsificação facial no mundo real.*Anais da Conferência IEEE/CVF sobre Visão Computacional e Reconhecimento de Padrões*, 2889–2898.

Karras, T., Laine, S., & Aila, T. (2019). Uma arquitetura de gerador baseada em estilo para redes adversárias generativas.*Anais da Conferência IEEE/CVF sobre Visão Computacional e Reconhecimento de Padrões*, 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analisando e melhorando a qualidade da imagem do StyleGAN. *Anais da Conferência IEEE/CVF sobre Visão Computacional e Reconhecimento de Padrões*, 8110–8119.

Kingston, J., Shaikh, S. A., & Serban, R. (2023). IA e fraude financeira: oportunidades, desafios e contramedidas.*Jornal de Crimes Financeiros*, 30(1), 21–35.

Kontakos, N., & Lioce, P. (2022). Fraude de CEO: combatendo golpes de clonagem de voz.*Revista de Cibersegurança*, 8(1), tyac019.

Kshetri, N. (2021). Inteligência artificial na detecção de fraudes financeiras: situação atual e direções futuras.*Jornal de Crimes Financeiros*, 28(1), 106–122.

Li, Y., Chang, MC e Lyu, S. (2020). In ictu oculi: Expondo vídeos falsos criados por IA detectando piscar de olhos. *Transações IEEE sobre Forense e Segurança da Informação*, 15, 2056–2068.

Matern, F., Riess, C., & Stamminger, M. (2019). Explorando artefatos visuais para expor deepfakes e manipulações faciais.*Anais da Conferência de Inverno do IEEE sobre Aplicações de Visão Computacional (WACV)*, 83–92.

Mirsky, Y., & Lee, W. (2021). A criação e detecção de deepfakes: uma pesquisa.*Pesquisas de Computação ACM (CSUR)*, 54(1), Artigo 14.

Navetta, D., Berglas, S., & Rotenberg, M. (2020). Inteligência artificial e cibercrime: Respostas jurídicas e práticas.*Revista Harvard de Direito e Tecnologia*, 33(2), 453–494.

Neekhara, P., Hussain, S., Pandey, P., Dubnov, S., Koushanfar, F., & McAuley, J. (2021). Ameaças adversas à detecção de deepfakes: uma perspectiva prática.*Anais da Conferência IEEE/CVF sobre Visão Computacional e Reconhecimento de Padrões (CVPR)*, 1561–1570.

Laboratório NTech. (2021).*Detecção de áudio deepfake: métodos e tecnologias*. Recuperado de https://ntechlab.com

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Aprendendo a detectar imagens faciais manipuladas.*Anais da Conferência Internacional IEEE sobre Visão Computacional (ICCV)*, 1–11.

Thies, J., Zollhöfer, M., & Nießner, M. (2019). Renderização neural diferida: síntese de imagens usando texturas neurais.*Transações ACM em Gráficos (TOG)*, 38(4), Artigo 66.

Tolosana, R., Romero-Tapiador, S., Fierrez, J., & Vera-Rodriguez, R. (2020). DeepFakes e além: uma pesquisa sobre manipulação facial e métodos de detecção de falsificações.*Fusão de informações*, 64, 131–148.

Tongco, M. D. C. (2007). Amostragem proposital como ferramenta para seleção de informantes.*Pesquisa e Aplicações em Etnobotânica*, 5, 147–158.

Verdoliva, L. (2020). Perícia de mídia e deepfakes: uma visão geral.*Revista IEEE de Tópicos Selecionados em Processamento de Sinais*, 14(5), 910–932.

Westerlund, M. (2019). O surgimento da tecnologia deepfake: uma revisão.*Revisão de Gestão da Inovação Tecnológica*, 9(11), 40–53.

Zhou, W., & Leung, K. S. (2021). Tecnologia financeira e abuso criminal: uma revisão e agenda de pesquisa.*Revista Internacional de Estudos Financeiros*, 9(1), 12.

LISTA DE TABELAS

Tabela 1: Incidentes relatados de fraude financeira habilitada por deepfake por ano (2019–2024).

Tabela 2: Tipos de tecnologia deepfake empregadas em incidentes de fraude.

Tabela 4: Modus operandi comum utilizado em fraudes financeiras habilitadas por deepfake.

Tabela 5: Distribuição setorial/industrial das entidades alvo.

Tabela 6: Métodos de detecção e resultados de eficácia