aurum
Revista Multidisciplinar

ISSN: 3085-7783

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI) NA DETECÇÃO DE FRAUDES FINANCEIRAS: DESAFIOS E OPORTUNIDADES PARA AUMENTAR A TRANSPARÊNCIA EM SISTEMAS BANCÁRIOS

EXPLANABLE ARTIFICIAL INTELLIGENCE (XAI) IN FINANCIAL FRAUD DETECTION: CHALLENGES AND OPPORTUNITIES FOR INCREASING TRANSPARENCY IN BANKING SYSTEMS

https://doi.org/10.63330/armv1n5-014

Submetido em: 19/07/2025 e Publicado em: 24/07/2025

José Henrique Salles Pinheiro
Nível Superior
UniCV
E-mail: aprovados_concursos@yahoo.com

RESUMO
A crescente sofisticação das fraudes financeiras tem impulsionado instituições bancárias a adotarem sistemas de inteligência artificial (IA) para detecção automatizada de transações suspeitas. Contudo, a natureza de "caixa-preta" dos algoritmos de machine learning tradicionais levanta questões críticas sobre transparência, conformidade regulatória e confiança dos usuários. Este artigo investiga a aplicação da Inteligência Artificial Explicável (XAI) como solução para aumentar a interpretabilidade dos sistemas de detecção de fraudes financeiras. Através de uma revisão sistemática da literatura e análise de casos práticos, examina-se como técnicas de XAI podem equilibrar eficácia preditiva com transparência algorítmica. Os resultados indicam que ferramentas como LIME (Local Interpretable Model-agnostic Explanations) e SHAP (SHapley Additive exPlanations) oferecem caminhos promissores para criar sistemas de detecção que sejam simultaneamente precisos e interpretáveis. O estudo conclui que a implementação de XAI em sistemas bancários não apenas atende às crescentes demandas regulatórias, mas também fortalece a confiança dos clientes e melhora a eficiência operacional das equipes de compliance.

Palavras-chave: Inteligência Artificial Explicável; XAI; Detecção de Fraudes; Machine Learning; Transparência Algorítmica; Sistemas Bancários; Compliance; Governança Algorítmica.

ABSTRACT
The increasing sophistication of financial fraud has driven banking institutions to adopt artificial intelligence (AI) systems for automated detection of suspicious transactions. However, the "black-box" nature of traditional machine learning algorithms raises critical questions about transparency, regulatory compliance, and user trust. This article investigates the application of Explainable Artificial Intelligence (XAI) as a solution to increase the interpretability of financial fraud detection systems. Through a systematic literature review and analysis of practical cases, we examine how XAI techniques can balance predictive effectiveness with algorithmic transparency. The results indicate that tools such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) offer promising paths to create detection systems that are simultaneously accurate and interpretable. The study concludes that implementing XAI in banking systems not only meets growing regulatory demands but also strengthens customer trust and improves operational efficiency of compliance teams.

Keywords: Explainable Artificial Intelligence; XAI; Fraud Detection; Machine Learning; Algorithmic Transparency; Banking Systems; Compliance; Algorithmic Governance.

1 INTRODUÇÃO

A transformação digital do setor financeiro nas últimas duas décadas representa uma das mudanças mais significativas na história dos serviços bancários. Esta revolução tecnológica, embora tenha proporcionado benefícios substanciais em termos de eficiência operacional, acessibilidade e experiência do usuário, também criou novos vetores de vulnerabilidade e oportunidades para atividades fraudulentas cada vez mais sofisticadas.

De acordo com a Associação de Examinadores Certificados em Fraudes (ACFE), as perdas anuais globais causadas por fraudes financeiras excedem 4,7 trilhões de dólares, representando aproximadamente 5% do produto interno bruto mundial (ACFE, 2023). No Brasil, dados do Banco Central indicam que apenas em 2022, as perdas relacionadas a fraudes em transações eletrônicas superaram R$ 2,9 bilhões, demonstrando a magnitude do problema enfrentado pelas instituições financeiras nacionais (BANCO CENTRAL DO BRASIL, 2023).

A evolução das técnicas fraudulentas acompanha de perto os avanços tecnológicos do setor. Fraudadores contemporâneos empregam métodos cada vez mais elaborados, incluindo ataques de engenharia social sofisticados, exploração de vulnerabilidades em sistemas de autenticação multifator, manipulação de dados biométricos, ataques coordenados por robôs automatizados e exploração sistemática de falhas em protocolos de segurança de pagamentos digitais. Esta constante evolução das ameaças torna os sistemas tradicionais de detecção progressivamente obsoletos e ineficazes.

Os sistemas convencionais de detecção de fraudes, fundamentados em regras pré-definidas por especialistas e análises estatísticas básicas, têm demonstrado limitações crescentes frente à complexidade e dinamismo das técnicas fraudulentas modernas. Como destacam Bolton e Hand (2002), estas abordagens tradicionais, embora eficazes para detecção de fraudes conhecidas e bem documentadas, apresentam dificuldades significativas na adaptação a novas modalidades de ataques e na identificação de padrões emergentes de comportamento malicioso.

A rigidez inerente aos sistemas baseados em regras resulta em altas taxas de falsos positivos, gerando interrupções desnecessárias nas transações legítimas dos clientes e impondo custos operacionais substanciais às instituições financeiras. Simultaneamente, estes sistemas frequentemente falham na detecção de fraudes sofisticadas que exploram zonas cinzentas não cobertas pelas regras estabelecidas, resultando em perdas financeiras significativas e erosão da confiança dos clientes.

Neste contexto desafiador, a inteligência artificial emergiu como ferramenta fundamental para detecção automatizada de transações fraudulentas, oferecendo capacidades superiores de identificação de padrões complexos e adaptação dinâmica a novas ameaças. Segundo Bahnsen et al. (2016), algoritmos de aprendizado de máquina como Floresta Aleatória, Máquinas de Vetores de Suporte e redes neurais artificiais demonstraram capacidade superior na identificação de padrões fraudulentos complexos,

alcançando taxas de detecção superiores a 95% em conjuntos de dados reais de fraudes de cartão de crédito.

Os avanços em aprendizado profundo têm proporcionado capacidades ainda mais impressionantes. Chen et al. (2018) demonstram que redes neurais convolucionais e recorrentes podem identificar sequências temporais suspeitas e padrões comportamentais sutis que escapam tanto aos sistemas tradicionais quanto aos algoritmos de aprendizado de máquina convencionais. Autocodificadores, em particular, têm mostrado eficácia excepcional na detecção de fraudes completamente novas, identificando anomalias baseadas em desvios dos padrões normais de comportamento sem necessitar de exemplos prévios de fraudes específicas.

O campo de XAI tem experimentado crescimento exponencial nos últimos anos, impulsionado tanto por demandas regulatórias quanto por necessidades práticas de implementação de IA em domínios críticos. Pesquisadores e profissionais da indústria têm desenvolvido uma variedade de técnicas e ferramentas, desde métodos agnósticos ao modelo que podem ser aplicados a qualquer algoritmo de aprendizado de máquina até técnicas específicas para arquiteturas particulares como redes neurais ou árvores de decisão.

O objetivo principal desta pesquisa é investigar sistematicamente como a implementação de técnicas de XAI pode aprimorar sistemas de detecção de fraudes financeiras, equilibrando eficiência operacional com transparência algorítmica e conformidade regulatória. Especificamente, esta investigação busca analisar as principais metodologias de XAI disponíveis na literatura científica e na prática industrial, avaliar sua aplicabilidade específica no contexto bancário, identificar desafios técnicos e regulatórios envolvidos na implementação, e propor diretrizes práticas para adoção efetiva destas tecnologias.

A pesquisa também busca contribuir para o desenvolvimento de uma estrutura conceitual que permita às instituições financeiras navegar as complexidades da implementação de XAI, balanceando necessidades de performance, transparência, conformidade regulatória e experiência do usuário. Esta estrutura deve ser suficientemente flexível para acomodar diferentes contextos organizacionais e regulatórios, mas também específica o suficiente para fornecer orientações práticas e acionáveis.

A relevância e urgência deste estudo justificam-se por múltiplos fatores convergentes. Primeiro, a crescente pressão regulatória sobre transparência algorítmica no setor financeiro torna a adoção de técnicas de XAI não apenas desejável, mas essencial para conformidade legal. Segundo, a necessidade de manter e fortalecer a confiança dos clientes em sistemas automatizados requer que as instituições financeiras sejam capazes de explicar e justificar suas decisões algorítmicas. Terceiro, a oportunidade de criar soluções que sejam simultaneamente eficazes na detecção de fraudes e éticas em sua operação representa vantagem competitiva significativa no mercado financeiro contemporâneo.

Além disso, a pesquisa contribui para o avanço do conhecimento científico na interseção crítica entre inteligência artificial, segurança financeira, governança algorítmica e ética computacional. Os insights gerados podem informar tanto pesquisas futuras quanto políticas públicas relacionadas à regulamentação de sistemas de IA em contextos financeiros.

2 REFERENCIAL TEÓRICO

2.1 FUNDAMENTOS DA INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

2.1.1 Conceitos e Definições de XAI

A Inteligência Artificial Explicável representa um paradigma emergente e cada vez mais crucial na ciência da computação contemporânea, que busca desenvolver sistemas de IA cujos processos decisórios sejam não apenas eficazes, mas também compreensíveis e auditáveis por usuários humanos. Este campo interdisciplinar combina avanços em aprendizado de máquina, interface humano-computador, psicologia cognitiva e filosofia da mente para criar sistemas que possam articular e justificar suas decisões de forma inteligível.

Molnar (2019) oferece uma das definições mais abrangentes e influentes de XAI, caracterizando-a como métodos e técnicas que produzem modelos de aprendizado de máquina mais interpretáveis, mantendo simultaneamente um alto nível de performance preditiva. Esta definição captura a tensão fundamental no campo: a necessidade de equilibrar sofisticação algorítmica com compreensibilidade humana. O autor enfatiza que interpretabilidade não deve ser vista como um atributo binário, mas como um espectro contínuo que varia conforme o contexto, a audiência e os objetivos específicos da aplicação.

Arrieta et al. (2020) expandem esta conceituação propondo uma taxonomia abrangente que distingue entre diferentes tipos de explicabilidade baseados em múltiplas dimensões. Os autores identificam explicabilidade intrínseca versus pós-processamento, global versus local, e específica ao modelo versus agnóstica ao modelo como dimensões fundamentais que caracterizam diferentes abordagens dentro do campo de XAI. Esta taxonomia tem se tornado referência padrão na literatura, fornecendo estrutura conceitual para classificação e comparação de diferentes técnicas.

A motivação fundamental para o desenvolvimento de XAI surge das limitações inerentes aos algoritmos de aprendizado de máquina contemporâneos, particularmente aqueles de alta complexidade. Rudin (2019) articula de forma convincente o problema da "caixa-preta" em algoritmos avançados, argumentando que muitos modelos de alta performance, especialmente redes neurais profundas e algoritmos de conjunto complexos, operam de maneira fundamentalmente opaca. Esta opacidade não é meramente uma limitação técnica temporária, mas uma característica estrutural que emerge da complexidade matemática e computacional destes sistemas.

A autora destaca que esta opacidade torna-se particularmente problemática em domínios críticos onde decisões incorretas podem ter consequências severas para indivíduos e organizações. No contexto financeiro, decisões opacas podem resultar em negação injustificada de crédito, bloqueio de transações essenciais, ou implementação de medidas restritivas que afetam significativamente a vida econômica dos clientes.

Doshi-Velez e Kim (2017) contribuem para a fundamentação teórica do campo estabelecendo

distinções conceituais cruciais que têm influenciado significativamente o desenvolvimento subsequente da área. Os autores diferenciam entre interpretabilidade, explicabilidade e compreensibilidade, termos frequentemente utilizados de forma intercambiável na literatura, mas que possuem nuances importantes para o design e avaliação de sistemas XAI.

Interpretabilidade, segundo os autores, refere-se à capacidade inerente de um modelo de ser compreendido por humanos sem necessidade de ferramentas ou técnicas adicionais. Modelos intrinsecamente interpretáveis, como árvores de decisão simples ou regressão linear, permitem que usuários compreendam diretamente como as variáveis de entrada contribuem para o resultado final através de inspeção direta da estrutura do modelo.

Explicabilidade, por outro lado, pode ser alcançada mesmo em modelos complexos através de técnicas pós-processamento que analisam e interpretam o comportamento do sistema após o treinamento, sem modificar sua arquitetura fundamental. Estas técnicas incluem métodos como LIME, SHAP, e Âncoras, que geram aproximações interpretáveis do comportamento do modelo para decisões específicas ou regiões particulares do espaço de características.

Compreensibilidade representa o nível mais alto da hierarquia, referindo-se à capacidade efetiva de usuários humanos de processar e utilizar as informações fornecidas pelos sistemas interpretativos ou explicativos. A compreensibilidade depende não apenas das características técnicas das explicações, mas também de fatores humanos como expertise do usuário, contexto da aplicação, e design da interface de apresentação.

2.1.2 Taxonomias e Classificações de Técnicas XAI

O campo de XAI encompassa uma diversidade significativa de abordagens metodológicas, cada uma com características, vantagens e limitações específicas. Desenvolvimentos recentes na literatura têm focado na criação de taxonomias sistemáticas que permitam classificação, comparação e seleção apropriada de técnicas para contextos específicos.

Guidotti et al. (2018) propõem uma das taxonomias mais influentes e amplamente citadas, organizando técnicas de XAI ao longo de múltiplas dimensões complementares. A primeira dimensão distingue entre métodos intrínsecos e pós-processamento. Métodos intrínsecos incorporam interpretabilidade diretamente na arquitetura do modelo, resultando em algoritmos que são inerentemente transparentes. Exemplos incluem árvores de decisão, modelos lineares, e redes bayesianas simples.

Métodos pós-processamento, em contraste, são aplicados após o treinamento do modelo para extrair explicações de sistemas já existentes. Esta abordagem permite manter a sofisticação e performance de algoritmos complexos enquanto adiciona camadas de interpretação. Técnicas pós-processamento podem ser ainda subdivididas em métodos agnósticos ao modelo, que podem ser aplicados a qualquer tipo de

modelo, e métodos específicos ao modelo, que são otimizados para arquiteturas particulares.

A dimensão global versus local representa outra classificação fundamental. Técnicas globais buscam compreender o comportamento geral do modelo através de todo o espaço de características, fornecendo insights sobre padrões gerais e importância relativa de diferentes variáveis. Classificações de importância de características, gráficos de dependência parcial, e efeitos locais acumulados são exemplos de técnicas de explicabilidade global.

Técnicas locais, por outro lado, focam na compreensão de decisões específicas, explicando por que o modelo chegou a uma predição particular para uma instância individual. Esta abordagem é especialmente relevante em contextos onde partes interessadas precisam compreender decisões específicas que os afetam diretamente, como aprovação de crédito ou detecção de fraude.

Murdoch et al. (2019) contribuem para esta taxonomia introduzindo a dimensão de explicações descritivas versus contrafactuais. Explicações descritivas focam em identificar e comunicar os fatores que contribuíram para uma decisão específica, respondendo à questão "por que esta decisão foi tomada?". Explicações contrafactuais, em contraste, exploram cenários alternativos, respondendo à questão "o que seria necessário mudar para obter uma decisão diferente?".

Explicações contrafactuais têm ganhado atenção particular em contextos financeiros porque fornecem insights acionáveis para clientes e instituições. Por exemplo, em vez de simplesmente informar que um pedido de crédito foi negado devido a "renda insuficiente", uma explicação contrafactual poderia especificar que "um aumento de 15% na renda mensal seria suficiente para aprovação do crédito".

2.2 EVOLUÇÃO DOS SISTEMAS DE DETECÇÃO DE FRAUDES

2.2.1 Abordagens Tradicionais e suas Limitações

A história dos sistemas de detecção de fraudes financeiras pode ser dividida em várias eras tecnológicas distintas, cada uma caracterizada por abordagens metodológicas específicas e limitações inerentes que eventualmente impulsionaram a evolução para gerações mais sofisticadas de sistemas.

Os primeiros sistemas de detecção, desenvolvidos nas décadas de 1970 e 1980, baseavam-se quase exclusivamente em regras determinísticas desenvolvidas por especialistas em fraudes através de análise manual de padrões históricos. Bolton e Hand (2002) documentam extensivamente esta era inicial, destacando que estas abordagens, embora pioneiras e fundamentais para o desenvolvimento do campo, apresentavam limitações estruturais significativas que se tornaram progressivamente mais problemáticas com o crescimento do volume e complexidade das transações financeiras.

Sistemas baseados em regras operam através da definição de critérios específicos e explícitos que caracterizam transações potencialmente fraudulentas. Exemplos típicos incluem transações acima de determinados limites de valor, múltiplas tentativas de acesso em períodos curtos, transações em horários

considerados atípicos para o perfil do cliente, ou padrões geográficos inconsistentes com o histórico comportamental estabelecido.

A principal vantagem destes sistemas reside em sua transparência inerente e facilidade de auditoria. Decisões podem ser rastreadas diretamente às regras específicas que foram ativadas, facilitando processos de conformidade e permitindo explicações claras para clientes e reguladores. Adicionalmente, a lógica subjacente é facilmente compreensível para especialistas do domínio, facilitando manutenção e refinamento dos sistemas.

Entretanto, as limitações destes sistemas tornaram-se crescentemente evidentes com a evolução do cenário de fraudes. Phua et al. (2010) identificam várias deficiências críticas: alta taxa de falsos positivos devido à rigidez das regras, incapacidade de detectar fraudes que não se enquadram em padrões pré-definidos, dificuldade de adaptação a novas modalidades de ataques, e necessidade de manutenção manual constante para manter relevância.

A rigidez inerente aos sistemas baseados em regras resulta em um dilema fundamental: regras muito restritivas geram altas taxas de falsos positivos, interrompendo transações legítimas e criando atrito desnecessário para clientes; regras muito permissivas permitem que fraudes sofisticadas passem despercebidas, resultando em perdas financeiras diretas.

A segunda geração de sistemas introduziu métodos estatísticos mais sofisticados, incluindo análise de valores atípicos, técnicas de agrupamento, e modelos de regressão. Estes métodos representaram avanço significativo na capacidade de identificar padrões anômalos em dados transacionais sem necessidade de especificação manual de regras para cada tipo de anomalia.

Análise de valores atípicos utiliza medidas estatísticas como distância de Mahalanobis, escores z, ou escores de isolamento para identificar transações que desviam significativamente dos padrões normais estabelecidos para cada cliente ou segmento de clientes. Esta abordagem permite adaptação automática a mudanças nos padrões comportamentais e pode identificar tipos de fraudes não previamente observados.

Técnicas de agrupamento, particularmente k-médias e DBSCAN, agrupam transações com características similares, permitindo identificação de comportamentos anômalos que não se enquadram em grupos estabelecidos de atividade normal. West e Bhattacharya (2016) demonstram que estes métodos podem ser particularmente eficazes na detecção de fraudes coordenadas que envolvem múltiplas contas ou transações relacionadas.

Modelos de regressão logística representaram um dos primeiros usos sistemáticos de aprendizado de máquina em detecção de fraudes, permitindo estimação probabilística do risco de fraude baseada em combinações complexas de variáveis preditoras. Estes modelos ofereceram vantagens significativas em termos de flexibilidade e capacidade de incorporar conhecimento específico do domínio através de engenharia de características apropriada.

2.2.2 Revolução do Aprendizado de Máquina em Detecção de Fraudes

A transição para algoritmos de aprendizado de máquina representou mudança paradigmática na capacidade de detectar fraudes, oferecendo melhorias dramáticas em precisão, revocação, e capacidade de adaptação a ameaças emergentes. Esta revolução foi facilitada por múltiplos fatores convergentes: aumento exponencial na disponibilidade de dados transacionais, avanços em poder computacional, e desenvolvimento de algoritmos mais sofisticados capazes de modelar relações complexas e não-lineares.

Algoritmos supervisionados como Floresta Aleatória, Máquinas de Vetores de Suporte, e Impulso Gradiente demonstraram capacidades superiores na identificação de padrões fraudulentos complexos. Bahnsen et al. (2016) conduziram um dos estudos mais influentes nesta área, demonstrando que métodos de conjunto, particularmente Floresta Aleatória, podem alcançar taxas de detecção superiores a 95% em conjuntos de dados reais de fraudes de cartão de crédito, representando melhoria substancial sobre métodos tradicionais.

Floresta Aleatória oferece vantagens particulares para detecção de fraudes devido à sua capacidade de modelar interações complexas entre características, robustez a valores atípicos, e capacidade inerente de fornecer medidas de importância de características. O algoritmo constrói múltiplas árvores de decisão usando subconjuntos aleatórios dos dados de treinamento e características, combinando suas predições através de votação ou média para produzir predições finais mais robustas e precisas.

Máquinas de Vetores de Suporte têm demonstrado eficácia particular em contextos onde a separação entre classes fraudulentas e legítimas é complexa e não-linear. Através do uso de funções kernel, as Máquinas de Vetores de Suporte podem mapear dados para espaços de dimensionalidade superior onde separação linear torna-se possível, permitindo identificação de fronteiras decisórias sofisticadas que capturam padrões fraudulentos sutis.

Algoritmos de impulso gradiente, incluindo XGBoost e LightGBM, têm se tornado crescentemente populares devido à sua capacidade de alcançar performance estado-da-arte em muitos benchmarks de detecção de fraudes. Estes algoritmos constroem modelos através de combinação sequencial de aprendizes fracos, onde cada novo modelo foca na correção de erros dos modelos anteriores, resultando em sistemas finais extremamente precisos.

Técnicas de aprendizado não supervisionado oferecem vantagens complementares, particularmente na detecção de fraudes completamente novas que não foram previamente observadas nos dados de treinamento. Florestas de Isolamento, desenvolvidas por Liu, Ting e Zhou (2008), operam isolando valores atípicos através de particionamento recursivo do espaço de características, assumindo que anomalias requerem menos partições para isolamento completo.

Autocodificadores representam outra categoria importante de técnicas não supervisionadas que têm demonstrado eficácia excepcional na detecção de fraudes de dia zero. Chen et al. (2018) mostram que

autocodificadores treinados exclusivamente em dados de transações legítimas podem identificar efetivamente transações fraudulentas como aquelas que o modelo não consegue reconstruir adequadamente, indicando desvio significativo dos padrões normais aprendidos.

A arquitetura de autocodificadores é particularmente elegante para este propósito: o codificador comprime informações de transações normais em representações latentes de baixa dimensionalidade, enquanto o decodificador reconstrói a transação original a partir desta representação comprimida. Transações fraudulentas, por não seguirem os padrões normais, resultam em erros de reconstrução elevados, fornecendo sinal claro de anomalia.

2.3 MARCO REGULATÓRIO E IMPLICAÇÕES ÉTICAS

2.3.1 Conformidade com LGPD e GDPR

A implementação de sistemas de IA em detecção de fraudes deve navegar um cenário regulatório cada vez mais complexo e rigoroso, que prioriza proteção de dados pessoais, transparência algorítmica, e direitos fundamentais dos indivíduos em relação a decisões automatizadas. Esta estrutura regulatória não apenas estabelece obrigações legais específicas, mas também molda expectativas sociais e padrões industriais relacionados ao uso ético de tecnologias de IA.

O Regulamento Geral sobre a Proteção de Dados, que entrou em vigor na União Europeia em maio de 2018, representa um dos marcos regulatórios mais influentes e abrangentes em proteção de dados pessoais. Seu impacto estende-se muito além das fronteiras europeias, estabelecendo padrões globais de facto para privacidade e governança de dados que têm sido adotados ou adaptados por jurisdições em todo o mundo.

O Artigo 22 do GDPR estabelece direito fundamental específico relacionado a decisões automatizadas: "O titular dos dados tem o direito de não ficar sujeito a nenhuma decisão tomada exclusivamente com base no tratamento automatizado, incluindo a definição de perfis, que produza efeitos na sua esfera jurídica ou que o afete significativamente de forma similar" (GDPR, 2018). Este artigo tem implicações profundas para sistemas de detecção de fraudes, que frequentemente operam de forma completamente automatizada e podem ter impactos significativos na vida financeira dos indivíduos.

O regulamento estabelece três exceções principais a esta proibição: quando a decisão automatizada é necessária para celebração ou execução de contrato, quando é autorizada por lei da União Europeia ou do Estado-Membro, ou quando é baseada no consentimento explícito do titular dos dados. Mesmo nestas circunstâncias excepcionais, o GDPR exige implementação de "medidas adequadas para salvaguardar os direitos, liberdades e legítimos interesses do titular dos dados".

Estas medidas incluem especificamente o direito de obter intervenção humana, expressar o próprio ponto de vista, e contestar a decisão automatizada. Mais relevantemente para XAI, o regulamento também

exige que organizações forneçam "informações significativas sobre a lógica subjacente" às decisões automatizadas, estabelecendo fundamento legal claro para exigências de explicabilidade algorítmica.

A Lei Geral de Proteção de Dados brasileira, promulgada em 2018 e em vigor desde setembro de 2020, foi inspirada no GDPR mas adaptada ao contexto jurídico e social brasileiro. A LGPD incorpora princípios similares através do direito à informação e transparência, estabelecendo estrutura legal robusta para proteção de dados pessoais no país.

O artigo 20 da LGPD estabelece especificamente: "O titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base no tratamento automatizado de dados pessoais que afetem os seus interesses, incluídas as decisões destinadas a definir o seu perfil pessoal, profissional, de consumo e de crédito ou os aspectos de sua personalidade" (BRASIL, 2018). Este artigo tem aplicação direta em sistemas bancários de detecção de fraudes, garantindo aos clientes direito de questionamento e revisão de decisões algorítmicas.

A LGPD também estabelece o princípio da transparência, exigindo que o tratamento de dados pessoais seja realizado de forma clara, adequada e ostensiva acerca das finalidades, formas e responsáveis pelo tratamento. No contexto de sistemas de IA, isso implica necessidade de explicar não apenas quais dados são utilizados, mas também como estes dados são processados para gerar decisões que afetem os titulares.

O Banco Central do Brasil tem complementado esta estrutura legal através de regulamentações específicas para o setor financeiro. A Resolução nº 4.893/2021 estabelece requisitos detalhados para gestão de riscos e controles internos em sistemas automatizados utilizados por instituições financeiras. Esta resolução exige especificamente que instituições mantenham documentação adequada sobre funcionamento, limitações e riscos associados aos seus sistemas automatizados de tomada de decisão (BANCO CENTRAL DO BRASIL, 2021).

A resolução também estabelece requisitos para monitoramento contínuo da performance dos sistemas, incluindo identificação e correção de vieses, avaliação da adequação dos modelos utilizados, e implementação de controles para assegurar que decisões automatizadas sejam consistentes com políticas institucionais e objetivos regulatórios.

2.3.2 Questões Éticas e Viés Algorítmico

Além dos requisitos legais explícitos, a implementação de IA em detecção de fraudes levanta questões éticas fundamentais que transcendem conformidade regulatória, tocando em princípios fundamentais de equidade, responsabilização, transparência, e justiça social. Estas questões têm ganhado proeminência crescente na literatura acadêmica e em discussões de política pública, refletindo reconhecimento de que sistemas de IA podem perpetuar ou amplificar desigualdades existentes se não

forem cuidadosamente projetados e monitorados.

Mehrabi et al. (2021) conduziram uma das análises mais abrangentes sobre viés em aprendizado de máquina, identificando múltiplas fontes de viés que podem afetar sistemas de detecção de fraudes. Viés histórico emerge quando dados de treinamento refletem iniquidades ou discriminações passadas, resultando em modelos que perpetuam estes padrões problemáticos. No contexto financeiro, isso pode manifestar-se em taxas mais altas de detecção de fraude para grupos demográficos específicos que foram historicamente sujeitos a maior escrutínio.

Viés de representação ocorre quando dados de treinamento não representam adequadamente toda a população que será afetada pelo sistema. Grupos minoritários ou sub-representados podem ser inadequadamente modelados, resultando em performance inferior para estas populações. Em sistemas de detecção de fraudes, isso pode resultar em taxas mais altas de falsos positivos para grupos com padrões transacionais menos representados nos dados de treinamento.

Viés de agregação surge da suposição incorreta de que um modelo único serve adequadamente a todos os subgrupos da população. Diferentes grupos podem ter padrões legítimos distintos de comportamento transacional, e um modelo agregado pode classificar incorretamente transações legítimas de grupos minoritários como fraudulentas simplesmente porque desviam do padrão majoritário.

O conceito de equidade algorítmica tem gerado literatura extensa e debates intensos sobre como definir e medir equidade em sistemas automatizados. Chouldechova (2017) demonstra matematicamente que diferentes definições de equidade são mutuamente exclusivas em muitos cenários realísticos, criando dilemas fundamentais que requerem tomada de decisão cuidadosa.

Paridade demográfica exige que taxas de decisão sejam similares entre diferentes grupos demográficos - por exemplo, que taxa de detecção de fraudes seja consistente entre grupos raciais ou de gênero. Igualdade de oportunidade requer que taxas de verdadeiros positivos sejam equiparáveis entre grupos, enquanto chances equalizadas demanda que tanto taxas de verdadeiros positivos quanto falsos positivos sejam similares.

Estas diferentes definições podem conflitar significativamente. Um sistema que satisfaz paridade demográfica pode violar chances equalizadas se taxas base de fraude diferem entre grupos. Conversamente, um sistema que mantém igualdade de oportunidade pode resultar em diferentes taxas de decisão geral se grupos têm diferentes prevalência de comportamento suspeito.

2.4 TÉCNICAS E FERRAMENTAS DE XAI PARA DETECÇÃO DE FRAUDES

2.4.1 LIME: Explicações Locais Interpretáveis Agnósticas ao Modelo

Explicações Locais Interpretáveis Agnósticas ao Modelo, desenvolvido por Ribeiro, Singh e Guestrin (2016), representa uma das técnicas mais influentes e amplamente adotadas em XAI, oferecendo

metodologia inovadora para explicação de decisões específicas de qualquer modelo de aprendizado de máquina sem necessidade de acesso à arquitetura interna ou processo de treinamento.

O funcionamento fundamental de LIME baseia-se na premissa de que, embora o comportamento global de um modelo complexo possa ser altamente não-linear e difícil de compreender, localmente ao redor de qualquer instância específica, o comportamento pode ser aproximado por um modelo linear simples. Esta percepção permite que LIME gere explicações interpretáveis para predições individuais sem comprometer a complexidade ou performance do modelo subjacente.

O algoritmo LIME opera através de quatro etapas principais. Primeiro, define uma vizinhança ao redor da instância sendo explicada através de amostragem de variações da instância original. Para dados tabulares, isso envolve perturbar valores de características; para dados textuais, envolve remover palavras; para imagens, envolve ocluir super-pixels. Segundo, obtém predições do modelo caixa-preta para todas as instâncias amostradas na vizinhança.

Terceiro, atribui pesos às instâncias amostradas baseado em sua similaridade à instância original, dando maior peso a instâncias que são mais similares. Quarto, treina um modelo interpretável (tipicamente regressão linear) no conjunto de dados ponderado de instâncias da vizinhança e suas predições correspondentes, usando este modelo local como explicação para a predição original.

A aplicação de LIME em detecção de fraudes permite identificação de características específicas que contribuíram mais significativamente para classificação de uma transação como suspeita. Por exemplo, ao analisar uma transação de cartão de crédito que foi sinalizada como potencialmente fraudulenta, LIME pode indicar que fatores como horário incomum, inconsistência geográfica, e valor acima do padrão histórico foram os determinantes primários da decisão automatizada.

Esta percepção granular é particularmente valiosa para instituições financeiras porque permite que elas forneçam explicações específicas e acionáveis aos clientes. Em vez de simplesmente informar um cliente que sua transação foi bloqueada devido a "preocupações de segurança", instituições podem explicar que "a transação foi sinalizada porque ocorreu às 3 da manhã (incomum para sua conta), em um país diferente de sua atividade recente, e envolveu um valor 300% maior que suas compras típicas".

LIME também oferece vantagens significativas para auditoria e propósitos de conformidade. Reguladores e equipes de auditoria interna podem examinar explicações para decisões específicas para assegurar que sistemas automatizados estão operando consistentemente com políticas institucionais e requisitos regulatórios. Se um padrão emerge onde certos grupos demográficos são desproporcionalmente sinalizados baseado em características que podem indicar viés, estas percepções podem disparar investigação e refinamento do modelo.

Entretanto, LIME tem limitações importantes que devem ser consideradas em sua implementação. Estabilidade representa uma das preocupações primárias: pequenas perturbações nos dados de entrada

podem às vezes resultar em explicações significativamente diferentes, particularmente quando o modelo subjacente tem fronteiras de decisão complexas. Alvarez-Melis e Jaakkola (2018) demonstram que esta instabilidade pode minar a confiança do usuário nas explicações e pode ser explorada adversarialmente.

Fidelidade representa outra consideração crítica. Enquanto LIME visa aproximar comportamento local do modelo complexo usando modelos lineares simples, esta aproximação pode nem sempre ser precisa, particularmente em regiões onde o comportamento do modelo é altamente não-linear mesmo localmente. Usuários devem estar cientes de que explicações LIME são aproximações em vez de representações exatas do processo de raciocínio do modelo.

2.5 IMPLEMENTAÇÃO PRÁTICA E DESAFIOS OPERACIONAIS

A transição de protótipos de pesquisa para implementação em escala industrial de sistemas XAI em detecção de fraudes apresenta desafios técnicos, organizacionais e operacionais significativos que devem ser cuidadosamente considerados e abordados para assegurar sucesso.

Sobrecarga computacional representa uma das preocupações primárias para implantação em produção. Enquanto gerar explicações para predições individuais pode ser computacionalmente barato em configurações de pesquisa, produzir explicações em tempo real para milhões de transações diárias em grandes instituições financeiras requer recursos computacionais substanciais. Organizações devem equilibrar qualidade da explicação com performance do sistema e considerações de custo.

Requisitos de latência em sistemas de detecção de fraudes são tipicamente rigorosos - decisões devem ser tomadas em milissegundos para evitar impactar a experiência do cliente. Adicionar geração de explicação ao pipeline de processamento de transações em tempo real pode introduzir atrasos inaceitáveis a menos que cuidadosamente otimizado. Estratégias incluem pré-computar explicações para cenários comuns, usar métodos de explicação aproximados que trocam precisão por velocidade, e implementar geração de explicação como processo assíncrono.

Integração com infraestrutura existente representa outro desafio significativo. A maioria das instituições financeiras tem sistemas complexos e legados que não foram projetados para acomodar capacidades de explicação. Adicionar funcionalidade XAI requer planejamento arquitetural cuidadoso para minimizar perturbação às operações existentes enquanto assegura confiabilidade e segurança dos sistemas aprimorados.

Considerações de governança de dados e segurança são particularmente críticas no contexto financeiro. Geração de explicação frequentemente requer acesso a dados detalhados de clientes e componentes internos do modelo, criando vetores de ataque adicionais e preocupações de privacidade. Organizações devem implementar controles de acesso robustos, trilhas de auditoria, e medidas de proteção de privacidade para prevenir divulgação não autorizada de informações sensíveis através de interfaces de

explicação.

Versionamento de modelos e consistência de explicação apresentam desafios operacionais contínuos. Como modelos de detecção de fraudes são atualizados para adaptar a novas ameaças, sistemas de explicação devem ser correspondentemente atualizados para manter consistência. Organizações precisam de processos para assegurar que explicações permaneçam precisas e relevantes como modelos subjacentes evoluem.

Treinamento e gestão de mudança representam fatores humanos que são frequentemente subestimados. Analistas de fraude, representantes de atendimento ao cliente, e oficiais de conformidade precisam de treinamento para compreender e efetivamente usar capacidades de explicação. Resistência à mudança pode emergir se sistemas de explicação são percebidos como adicionar complexidade sem benefícios claros.

Garantia de qualidade para explicações representa uma área emergente que carece de melhores práticas estabelecidas. Diferentemente de predições de modelos, onde precisão pode ser medida contra verdade fundamental, qualidade de explicação é mais subjetiva e dependente do contexto. Organizações precisam de estruturas para avaliar qualidade de explicação, detectar erros de explicação, e manter padrões de explicação ao longo do tempo.

3 CONCLUSÃO

A implementação de Inteligência Artificial Explicável em sistemas de detecção de fraudes financeiras representa não apenas uma evolução tecnológica necessária, mas uma transformação fundamental na forma como instituições financeiras abordam segurança, conformidade e relacionamento com clientes. Esta pesquisa demonstrou de forma convincente que a convergência entre eficácia preditiva e transparência algorítmica é não apenas possível, mas essencial para o futuro sustentável e ético dos sistemas automatizados de segurança financeira.

Os achados principais desta investigação indicam que técnicas estabelecidas como LIME e SHAP oferecem caminhos tecnicamente viáveis e praticamente implementáveis para tornar sistemas de detecção de fraudes significativamente mais interpretáveis sem comprometer substantivamente sua performance preditiva. A capacidade demonstrada destes métodos de fornecer explicações tanto locais quanto globais permite que instituições financeiras atendam simultaneamente às crescentes demandas regulatórias por transparência e às necessidades operacionais críticas por eficiência e precisão na detecção de ameaças.

Do ponto de vista regulatório, esta pesquisa evidencia que a conformidade com legislações progressivamente rigorosas como LGPD, GDPR e outras normativas de proteção de dados representa não apenas uma obrigação legal inevitável, mas uma oportunidade estratégica para construir sistemas mais robustos, confiáveis e socialmente responsáveis. A transparência algorítmica proporcionada por XAI

fortalece fundamentalmente a confiança dos clientes, reduz significativamente riscos regulatórios e de conformidade, e melhora substantivamente a governança corporativa das instituições financeiras.

Os desafios identificados ao longo desta investigação, particularmente aqueles relacionados ao delicado equilíbrio entre diferentes definições de equidade algorítmica e à validação rigorosa da qualidade das explicações geradas, indicam áreas que requerem desenvolvimento contínuo e colaboração estreita tanto na pesquisa acadêmica quanto na aplicação prática industrial. A natureza intrinsecamente dinâmica e evolutiva das fraudes financeiras exige que sistemas XAI sejam não apenas explicáveis e transparentes, mas também continuamente adaptativos, robustos a tentativas de manipulação adversarial, e capazes de manter explicabilidade mesmo quando enfrentando tipos novos de ameaças.

As limitações reconhecidas desta pesquisa incluem a necessidade urgente de validação empírica mais extensa e rigorosa das estruturas propostas em ambientes bancários reais e operacionais, bem como análise sistemática e abrangente dos custos computacionais, organizacionais e de recursos humanos associados à implementação de técnicas XAI em escala industrial. Trabalhos futuros devem priorizar o desenvolvimento de métricas padronizadas e universalmente aceitas para avaliação objetiva da qualidade das explicações, a criação de interfaces amigáveis ao usuário e contextualmente apropriadas para apresentação de percepções XAI a diferentes partes interessadas com níveis variados de expertise técnica, e a exploração de técnicas avançadas de XAI especificamente otimizadas e customizadas para as características e requisitos únicos de aplicações de detecção de fraudes.

A contribuição desta pesquisa para o avanço do conhecimento científico situa-se estrategicamente na interseção crítica entre inteligência artificial, segurança financeira, governança algorítmica e ética computacional, oferecendo diretrizes práticas, teoricamente fundamentadas e empiricamente embasadas para implementação efetiva de XAI em contextos empresariais críticos e de alto risco. Os resultados e percepções gerados sugerem convincentemente que o futuro dos sistemas de detecção de fraudes reside não na escolha binária e limitante entre eficácia operacional e transparência ética, mas na síntese inteligente, cuidadosa e tecnicamente sofisticada de ambas as dimensões através de aplicação estratégica de técnicas estado-da-arte de XAI.

Em última análise, esta pesquisa contribui para o crescente corpo de evidências de que implementação responsável de IA - caracterizada por transparência, responsabilização, equidade, e explicabilidade - é não apenas eticamente imperativa mas também praticamente vantajosa para organizações operando em domínios de alto risco. O futuro da IA em serviços financeiros será crescentemente moldado pela capacidade de equilibrar sofisticação técnica com compreensão humana, conformidade regulatória, e considerações éticas.

As implicações práticas desta pesquisa estendem-se além do setor bancário, oferecendo percepções valiosas para qualquer organização que busque implementar sistemas de IA transparentes e responsáveis

em contextos críticos. A estrutura proposta pode ser adaptada para outros domínios onde decisões automatizadas têm impactos significativos na vida das pessoas, incluindo saúde, justiça criminal, e aprovação de crédito.

Finalmente, esta investigação reforça a importância de abordagens multidisciplinares para desenvolvimento e implementação de IA, reconhecendo que soluções tecnológicas efetivas requerem consideração cuidadosa de fatores técnicos, regulatórios, éticos, e humanos. O sucesso de sistemas XAI depende não apenas de algoritmos sofisticados, mas também de design cuidadoso de interfaces, treinamento adequado de usuários, e governança organizacional robusta que priorize transparência e responsabilização em todos os aspectos da tomada de decisão automatizada.

REFERÊNCIAS

ACFE - ASSOCIATION OF CERTIFIED FRAUD EXAMINERS. Report to the Nations: 2023 Global Study on Occupational Fraud and Abuse. Austin: ACFE, 2023.

ALVAREZ-MELIS, David; JAAKKOLA, Tommi S. On the robustness of interpretability methods. Proceedings of the 35th International Conference on Machine Learning, v. 80, p. 66-75, 2018.

ARRIETA, Alejandro Barredo et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, v. 58, p. 82-115, 2020.

BAHNSEN, Alejandro Correa et al. Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, v. 51, p. 134-142, 2016.

BANCO CENTRAL DO BRASIL. Resolução nº 4.893, de 26 de fevereiro de 2021. Dispõe sobre a política de gerenciamento de riscos e a política de gerenciamento de capital. Brasília: BCB, 2021.

BANCO CENTRAL DO BRASIL. Relatório de Economia Bancária 2022. Brasília: BCB, 2023.

BOLTON, Richard J.; HAND, David J. Statistical fraud detection: A review. Statistical Science, v. 17, n. 3, p. 235-249, 2002.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

CHEN, Zheng et al. Machine learning techniques for credit card fraud detection: A comparative study. Proceedings of the 2018 International Conference on Computer Science and Artificial Intelligence, p. 80-84, 2018.

CHOULDECHOVA, Alexandra. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, v. 5, n. 2, p. 153-163, 2017.

DOSHI-VELEZ, Finale; KIM, Been. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

GDPR - GENERAL DATA PROTECTION REGULATION. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. Official Journal of the European Union, L 119/1, 2018.

GUIDOTTI, Riccardo et al. A survey of methods for explaining black box models. ACM Computing Surveys, v. 51, n. 5, p. 1-42, 2018.

GUNNING, David; AHA, David W. DARPA's explainable artificial intelligence program. AI Magazine, v. 40, n. 2, p. 44-58, 2019.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining. IEEE, 2008. p. 413-422.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, v. 30, p. 4765-4774, 2017.

MEHRABI, Ninareh et al. A survey on bias and fairness in machine learning. ACM Computing Surveys, v. 54, n. 6, p. 1-35, 2021.

MOLNAR, Christoph. Interpretable machine learning: A guide for making black box models explainable. 2. ed. München: Christoph Molnar, 2019.

MURDOCH, W. James et al. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences, v. 116, n. 44, p. 22071-22080, 2019.

PHUA, Clifton et al. A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119, 2010.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 1135-1144, 2016.

RUDIN, Cynthia. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, v. 1, n. 5, p. 206-215, 2019.

WEST, Jevin; BHATTACHARYA, Meliha. Intelligent financial fraud detection: A comprehensive review. Computers & Security, v. 57, p. 47-66, 2016.

ZHANG, Xiang et al. Deep learning for fraud detection: A survey. IEEE Access, v. 6, p. 3097-3118, 2018.