www.revistaft.com.br

Revista FT | ISSN 1678-0817 | v. 30, n. 164, 2026 | DOI: 10.69849/6va6aj45

# Detecção de fraudes em transações financeiras com *Machine Learning*: uma revisão da literatura

Fraud detection in financial transactions with *Machine Learning*: a literature review

Alexandre Siqueira Souza Costa[1]
João Paulo de Assis Marques[2]
Fabio Almeida do Nascimento[3]

**RESUMO**

A crescente digitalização das transações financeiras ampliou a ocorrência de fraudes, tornando sua detecção um desafio central para instituições bancárias e empresas de meios de pagamento. Diante da limitação dos sistemas tradicionais baseados em regras fixas, o aprendizado de máquina (*machine learning*) passou a ocupar posição de destaque na literatura científica recente. Este estudo teve como objetivo analisar, por meio de revisão bibliográfica narrativa, as principais técnicas de *machine learning* aplicadas à detecção de fraudes em transações financeiras, com destaque para cartões de crédito. A pesquisa, de natureza qualitativa e exploratória, baseou-se na análise crítica de oito trabalhos científicos publicados entre 2018 e 2026, localizados em bases como IEEE Xplore, Scopus, ScienceDirect e Google Scholar. Os resultados indicam o uso recorrente de algoritmos de *ensemble*, como Random Forest e *boosting*, ao lado de modelos clássicos, como Regressão Logística, Árvores de Decisão e Support Vector Machine, sem que haja um algoritmo universalmente superior. O desbalanceamento extremo dos dados é tratado por subamostragem, pelo SMOTE e suas variantes e pelo ajuste do limiar de decisão. Identificou-se ainda que métricas como F1-Score, AUC-ROC e área sob a curva de precisão-*recall* são mais adequadas do que a acurácia

[1] Fatec Praia Grande - E-mail: alexandre.costa5@aluno.cps.sp.gov.br
[2] Fatec de Praia Grande - E-mail: joão.marques3@aluno.cps.sp.gov.br
[3] Fatec Praia Grande - E-mail: fabio.nascimento@cps.sp.gov.br

isolada, e discutiu-se a tensão entre modelos opacos e o direito à revisão de decisões automatizadas previsto na Lei Geral de Proteção de Dados Pessoais. Conclui-se que persistem lacunas relacionadas à interpretabilidade dos modelos, à adaptação a fraudes emergentes e à escassez de dados reais e públicos para validação, apontando direções para pesquisas futuras.
**Palavras-Chave:** Aprendizado de máquina; Detecção de fraudes; Desbalanceamento de dados; Segurança da informação; Transações financeiras.

## ABSTRACT

The growing digitalization of financial transactions has significantly increased the occurrence of fraud, making its detection a central challenge for banking institutions and payment companies. Given the limitations of traditional rule-based systems, machine learning has gained a prominent position in recent scientific literature. This study aimed to analyze, through a narrative literature review, the main machine learning techniques applied to fraud detection in financial transactions, with emphasis on credit cards. The research, qualitative and exploratory in nature, was based on the critical analysis of eight scientific works published between 2018 and 2026, found in databases such as IEEE Xplore, Scopus, ScienceDirect, and Google Scholar. Results indicate the recurrent use of ensemble algorithms, such as Random Forest and boosting, alongside classical models, such as Logistic Regression, Decision Trees, and Support Vector Machine, with no universally superior algorithm. The extreme data imbalance is addressed through undersampling, SMOTE and its variants, and decision threshold adjustment. It was also found that metrics such as F1-Score, AUC-ROC, and the area under the precision-recall curve are more suitable than accuracy alone, and the tension between opaque models and the right to review automated decisions set out in the Brazilian General Data Protection Law was discussed. It is concluded that gaps remain regarding model interpretability, adaptation to emerging fraud patterns, and the scarcity of real and public data for validation, pointing to relevant directions for future research.
**KEYWORDS:** Machine learning; Fraud detection; Data imbalance; Information security; Financial transactions.

## 1 INTRODUÇÃO

A transformação digital dos serviços financeiros alterou profundamente a forma como pessoas e empresas realizam pagamentos, transferências e demais operações bancárias. Cartões de crédito, carteiras digitais e sistemas de pagamento instantâneo tornaram-se

ferramentas centrais do cotidiano econômico, ampliando o volume e a velocidade das transações processadas diariamente pelas instituições financeiras. Esse cenário, embora traga ganhos evidentes de conveniência e eficiência, também expandiu consideravelmente as oportunidades para a prática de fraudes financeiras.

A fraude em transações financeiras representa um problema de escala global, com impacto econômico estimado em US$416 bilhões em 2017 (RYMAN-TUBB; KRAUSE; GARN, 2018), afetando instituições bancárias, comerciantes e consumidores. Sistemas tradicionais de detecção, fundamentados em regras fixas definidas por especialistas, mostram-se cada vez mais insuficientes diante da sofisticação e da constante mutação das estratégias empregadas por fraudadores. Essa limitação decorre, em grande medida, da rigidez desses sistemas, que não conseguem se adaptar automaticamente a novos padrões de comportamento fraudulento sem intervenção manual constante.

Nesse contexto, o aprendizado de máquina (*machine learning*) emergiu como abordagem tecnológica capaz de superar parte dessas limitações, ao permitir que sistemas aprendessem padrões complexos diretamente dos dados históricos de transações, identificando anomalias e comportamentos suspeitos de forma adaptativa. Algoritmos supervisionados, não supervisionados e híbridos vêm sendo amplamente investigados na literatura científica como alternativas eficazes para a detecção de fraudes, apresentando resultados promissores em termos de precisão e capacidade de generalização (RYMAN-TUBB; KRAUSE; GARN, 2018).

Entretanto, a aplicação de *machine learning* à detecção de fraudes enfrenta desafios metodológicos relevantes. Um dos mais críticos é o desbalanceamento extremo característico desses conjuntos de dados, nos quais transações fraudulentas representam parcela muito pequena do total de registros: 0,172% na base pública de transações europeias utilizada por Salekshahrezaee, Leevy e Khoshgoftaar (2023) e por Oliveira (2025). Esse desequilíbrio compromete o desempenho de diversos algoritmos, que tendem a favorecer a classe majoritária, dificultando a identificação correta das transações fraudulentas. Some-se a isso a necessidade de avaliar os modelos por meio de métricas adequadas, uma vez que a acurácia isolada é insuficiente e, muitas vezes, enganosa nesse tipo de problema.

Diante desse cenário, formula-se o problema orientador desta pesquisa: quais técnicas de *machine learning* têm sido empregadas na literatura científica para a detecção de fraudes em transações financeiras, e de que forma os desafios relacionados ao desbalanceamento de dados, à avaliação de desempenho e às exigências legais de explicabilidade e privacidade têm sido enfrentados pelos pesquisadores da área?

O objetivo geral deste trabalho consiste em analisar, por meio de revisão bibliográfica, as principais técnicas de *machine learning* aplicadas à detecção de fraudes em transações financeiras. Para tanto, estabelecem-se os seguintes objetivos específicos: (a) identificar os algoritmos de *machine learning* mais utilizados na literatura para essa finalidade; (b) analisar as técnicas empregadas para o tratamento do desbalanceamento de classes; (c) levantar as métricas de avaliação de desempenho mais adequadas a esse tipo de problema; (d) discutir os principais desafios e lacunas apontados pelos autores, apontando direções para pesquisas futuras; e (e) examinar as implicações da Lei Geral de Proteção de Dados Pessoais (LGPD) para o uso de modelos de decisão automatizada na detecção de fraudes, em especial quanto ao direito à revisão dessas decisões, à informação sobre seus critérios e à privacidade.

A relevância deste estudo justifica-se pela importância crescente do tema tanto para a comunidade acadêmica quanto para o setor produtivo. Do ponto de vista acadêmico, a consolidação de uma síntese crítica da literatura recente contribui para organizar um campo de pesquisa disperso em múltiplas áreas de conhecimento (ciência da computação, estatística e segurança da informação). Do ponto de vista prático, os resultados podem orientar profissionais de tecnologia e instituições financeiras na escolha de abordagens tecnicamente mais robustas para a construção de sistemas antifraude, contribuindo indiretamente para a proteção de dados pessoais e financeiros dos usuários, tema também disciplinado pela Lei Geral de Proteção de Dados Pessoais no Brasil (BRASIL, 2018).

Este trabalho está organizado em cinco seções, incluindo esta introdução. A segunda seção apresenta o referencial teórico, abordando os conceitos de fraude financeira, os fundamentos do aprendizado de máquina aplicados a esse domínio e o desafio do desbalanceamento de classes. A terceira seção descreve a metodologia empregada na condução da revisão bibliográfica. A quarta seção expõe e discute os resultados obtidos a partir da análise dos estudos selecionados, incluindo uma discussão sobre decisão automatizada, explicabilidade e privacidade à luz da LGPD. Por fim, a quinta seção apresenta as considerações finais, as limitações do estudo e as recomendações para trabalhos futuros.

## 2 REFERENCIAL TEÓRICO

Este referencial teórico apresenta a estrutura conceitual que fundamenta a pesquisa, abrangendo os conceitos de fraude em transações financeiras, os fundamentos do aprendizado de máquina aplicados à sua detecção e o desafio do desbalanceamento de classes, recorrente nesse tipo de problema.

## 2.1 Fraude em transações financeiras: conceitos e contexto

A fraude em cartões de crédito pode ser definida como o uso não autorizado de informações de pagamento com o objetivo de obter bens, serviços ou valores financeiros de forma ilícita (RYMAN-TUBB; KRAUSE; GARN, 2018). Historicamente, a detecção desse tipo de fraude baseava-se em sistemas de regras estáticas, elaborados por especialistas de domínio a partir de padrões previamente conhecidos de comportamento fraudulento. Embora eficazes contra ameaças já catalogadas, esses sistemas demonstram baixa capacidade de adaptação frente a novas modalidades de fraude, exigindo atualização manual constante e apresentando elevadas taxas de falsos positivos (RYMAN-TUBB; KRAUSE; GARN, 2018).

Com o crescimento exponencial do volume de transações eletrônicas, tornou-se inviável depender exclusivamente da análise humana ou de regras fixas para identificar comportamentos anômalos em tempo hábil. Esse cenário impulsionou a adoção de técnicas de mineração de dados e, posteriormente, de aprendizado de máquina, capazes de processar grandes volumes de dados e identificar padrões sutis que escapariam à análise tradicional (RYMAN-TUBB; KRAUSE; GARN, 2018).

No Brasil, a digitalização acelerada dos meios de pagamento amplia a relevância de mecanismos automatizados de detecção, capazes de identificar atividades suspeitas em tempo real e de evitar alarmes falsos que geram insatisfação de clientes legítimos (TOSTA; DIAS, 2025). Adicionalmente, o tratamento dos dados pessoais utilizados nesses sistemas deve observar os princípios estabelecidos pela Lei Geral de Proteção de Dados Pessoais, entre eles os da finalidade, da necessidade e da segurança (BRASIL, 2018).

## 2.2 Fundamentos do aprendizado de máquina aplicado à detecção de fraudes

O aprendizado de máquina compreende um conjunto de técnicas computacionais que permitem a um sistema aprender padrões a partir de dados, sem que todas as regras de decisão sejam explicitamente programadas. Na detecção de fraudes, essa capacidade é particularmente valiosa, pois os padrões fraudulentos evoluem continuamente, tornando inviável a manutenção de regras estáticas atualizadas manualmente (RYMAN-TUBB; KRAUSE; GARN, 2018).

As abordagens supervisionadas constituem a categoria predominante nos trabalhos analisados, uma vez que pressupõem bases com rótulos indicando se cada transação é legítima ou fraudulenta. Entre os algoritmos supervisionados recorrentes, destacam-se a Regressão Logística, o Naive Bayes, as Árvores de Decisão e o Support Vector Machine (SVM)

(OLIVEIRA, 2025), o Random Forest e as redes neurais (TOSTA; DIAS, 2025) e algoritmos de *boosting*, como o XGBoost, o CatBoost e o LightGBM (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023). Métodos de *ensemble*, que combinam múltiplos classificadores para produzir uma decisão mais robusta, apresentam desempenho sólido em diversos estudos, conforme registram Salekshahrezaee, Leevy e Khoshgoftaar (2023), e o Random Forest e o Gradient Boosting figuram entre os algoritmos de bom desempenho identificados na revisão de Penha, Moraes e Melo (2026).

Técnicas de aprendizado profundo (*deep learning*), como Redes Neurais Convolucionais (CNN) e Long Short-Term Memory (LSTM), também vêm ganhando espaço na literatura recente sobre detecção de fraudes em cartões (HAFEZ et al., 2025); as redes LSTM, em particular, são capazes de capturar relações temporais em sequências de transações. *Autoencoders* convolucionais também têm sido empregados como técnica de extração de atributos, sendo combinados a classificadores de *ensemble* e comparados a um método tradicional de redução de dimensionalidade, a Análise de Componentes Principais (PCA) (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023). Apesar do desempenho elevado, esses modelos demandam maior volume de dados e poder computacional, além de apresentarem menor interpretabilidade em comparação a métodos mais tradicionais.

A construção de um modelo de detecção de fraudes envolve, tipicamente, etapas de pré-processamento dos dados (limpeza, normalização e engenharia de atributos), seleção e treinamento do algoritmo, validação cruzada e, por fim, avaliação de desempenho por meio de métricas específicas. Cada uma dessas etapas influencia diretamente a capacidade do modelo de identificar fraudes reais sem comprometer excessivamente a experiência de usuários legítimos (RYMAN-TUBB; KRAUSE; GARN, 2018).

### 2.3 O desafio do desbalanceamento de classes

Um dos aspectos mais discutidos na literatura sobre detecção de fraudes é o desbalanceamento extremo entre as classes de interesse. Na base pública de transações europeias de setembro de 2013, por exemplo, há 492 fraudes em 284.807 transações, ou 0,172% do total (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023; OLIVEIRA, 2025), o que caracteriza um cenário de desbalanceamento extremo. Esse desequilíbrio compromete diretamente o desempenho de algoritmos de classificação tradicionais, que tendem a otimizar a acurácia geral em detrimento da correta identificação da classe minoritária, isto é, das próprias fraudes.

Para mitigar esse problema, diversas técnicas de balanceamento de dados têm sido propostas. O Synthetic Minority Over-sampling Technique (SMOTE), proposto originalmente por Chawla et al. (2002), consiste na geração de amostras sintéticas da classe minoritária a partir da interpolação entre exemplos existentes, ampliando artificialmente sua representatividade sem simplesmente duplicar registros. A técnica é avaliada, por exemplo, por Salekshahrezaee, Leevy e Khoshgoftaar (2023), em combinação com os classificadores Random Forest, CatBoost, LightGBM e XGBoost.

Além do SMOTE, outras estratégias incluem a subamostragem aleatória (*random undersampling* – RUS) da classe majoritária e técnicas híbridas, como o SMOTE-Tomek, que combina a sobreamostragem com a remoção dos chamados pares de Tomek, isto é, pares de exemplos de classes diferentes em que um deles é ruído ou ambos estão na fronteira entre as classes (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023). Cada técnica apresenta limitações: a subamostragem descarta exemplos da classe majoritária (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023), enquanto a sobreamostragem por simples replicação torna muito específica a região de decisão da classe minoritária, o que equivale a *overfitting*, problema que motivou a proposta do SMOTE (CHAWLA et al., 2002).

Complementarmente ao balanceamento dos dados, há abordagens no nível do algoritmo, tipicamente baseadas em aprendizado sensível ao custo (*cost-sensitive learning*) (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023), que atribuem penalidades diferenciadas aos erros de classificação conforme a classe envolvida. Desse modo, prioriza-se a correta identificação das fraudes, ainda que à custa de um número maior de falsos positivos, cujo custo operacional tende a ser inferior ao de fraudes não detectadas.

## 3 METODOLOGIA

A presente pesquisa caracteriza-se como uma revisão bibliográfica narrativa, de natureza qualitativa e caráter exploratório, cujo objetivo é sintetizar e analisar criticamente a produção científica recente sobre a aplicação de *machine learning* à detecção de fraudes em transações financeiras. Optou-se pela revisão narrativa, em vez de uma revisão sistemática com protocolo rígido, em razão da natureza introdutória deste trabalho e do tempo disponível para sua elaboração, sem que isso comprometesse o rigor na seleção e na análise das fontes utilizadas.

O levantamento das fontes foi realizado nas bases IEEE Xplore, ScienceDirect, Scopus, SpringerLink e Google Scholar; os periódicos nacionais foram localizados

principalmente por meio do Google Scholar. As buscas foram conduzidas utilizando combinações dos descritores "fraud detection", "credit card fraud", "*machine learning*", "class imbalance" e seus equivalentes em português, como "detecção de fraudes", "aprendizado de máquina" e "desbalanceamento de dados".

Como critérios de inclusão, foram considerados artigos publicados entre 2018 e 2026, disponíveis em texto completo, redigidos em português, inglês ou espanhol, e diretamente relacionados à aplicação de técnicas de *machine learning* na detecção de fraudes em transações financeiras. Foram excluídos trabalhos sem relação direta com o tema, resumos sem acesso ao texto completo e publicações de caráter puramente comercial ou não revisadas por pares. Obras fundamentais para a compreensão de conceitos-base foram mantidas mesmo fora do intervalo temporal definido, dada sua relevância seminal: a técnica SMOTE (CHAWLA et al., 2002), a análise de curvas de precisão-*recall* em bases desbalanceadas (SAITO; REHMSMEIER, 2015) e as técnicas de explicabilidade LIME (RIBEIRO; SINGH; GUESTRIN, 2016) e SHAP (LUNDBERG; LEE, 2017). Essas obras, assim como a diretriz PRISMA (PAGE et al., 2021) e a legislação brasileira citada (BRASIL, 2018), foram utilizadas como fontes de apoio e não integram o corpus de análise.

A seleção das fontes não seguiu um protocolo sistemático de triagem. As buscas, realizadas nas bases e nos periódicos citados com os descritores indicados, resultaram em trabalhos escolhidos conforme sua relevância, sua pertinência ao tema e a disponibilidade em texto completo. Como não houve registro quantitativo das etapas de identificação e triagem, não é possível informar o total de registros retornados, o que constitui uma limitação deste estudo. O corpus final de análise foi composto por oito trabalhos científicos, sintetizados no Quadro 1 da seção 4.

A análise dos materiais selecionados seguiu três etapas: leitura exploratória dos títulos e resumos, para avaliar a aderência ao tema; leitura seletiva dos textos completos considerados pertinentes; e leitura analítica, voltada à extração organizada de informações relevantes, como algoritmos utilizados, técnicas de balanceamento de dados, métricas de avaliação empregadas e principais achados e limitações relatados pelos autores. As informações extraídas foram organizadas em categorias temáticas, que estruturam a apresentação dos resultados na seção seguinte: algoritmos mais utilizados, técnicas de balanceamento de dados, métricas de avaliação de desempenho, desafios e lacunas identificados na literatura e implicações da LGPD para a decisão automatizada.

## 4 RESULTADOS E DISCUSSÃO

A análise da literatura selecionada permitiu identificar padrões consistentes quanto às técnicas de *machine learning* empregadas na detecção de fraudes financeiras, bem como os principais desafios enfrentados pelos pesquisadores da área. O Quadro 1 sintetiza os trabalhos do corpus, e a discussão a seguir está organizada nos eixos temáticos que emergiram dessa categorização.

**Quadro 1 – Síntese dos trabalhos que compõem o corpus**

| Obra | Tipo | Algoritmos e técnicas | Tratamento do desbalanceamento | Métricas e base de dados |
|---|---|---|---|---|
| Ryman-Tubb, Krause e Garn (2018) | Revisão e *benchmark* da indústria | Métodos de IA e ML publicados; apenas oito com desempenho prático para a indústria | — | *Benchmark* proposto pelos autores |
| Dal Pozzolo et al. (2018) | Empírico | Classificadores baseados em Random Forest; estratégia para *concept drift* e atraso de rótulos | Subamostragem da classe majoritária | Precisão dos alertas e AUC; base real com mais de 75 milhões de transações |
| Salekshahrezaee, Leevy e Khoshgoftaar (2023) | Empírico | Random Forest, CatBoost, LightGBM e XGBoost; extração de atributos por PCA e CAE | RUS, SMOTE e SMOTE-Tomek | F1-Score e AUC; base europeia (0,172% de fraudes) |
| Souza e Bordin Jr. (2023) | Tutorial | Aprendizado supervisionado (métodos clássicos e híbridos) | Técnicas usuais de balanceamento | Métricas de classificação |
| Hafez et al. (2025) | Revisão sistemática | *Machine learning*, *deep learning* e meta-heurísticas | Apontado como limitação recorrente | Acurácia, precisão, sensibilidade e F1-Score (estudos revisados) |
| Oliveira (2025) | Empírico | Regressão Logística, Naive Bayes, SVM e Árvore de Decisão | Ajuste do limiar de decisão | Acurácia, precisão, *recall*, F1-Score e AUC; base europeia |
| Tosta e Dias (2025) | Revisão bibliográfica | Árvores de Decisão, SVM, Random Forest, redes neurais, k-means e Isolation Forest | — | — |
| Penha, Moraes e Melo (2026) | Revisão de literatura (12 estudos) | Random Forest e Gradient Boosting com bom desempenho; modelos híbridos | Técnicas de balanceamento | — |

Fonte: elaborado pelo autor com base nos trabalhos analisados (2026). Nota: "—" indica informação não especificada ou não aplicável.

### 4.1 Algoritmos de *machine learning* mais utilizados

Os estudos analisados evidenciam predominância de algoritmos de aprendizado supervisionado na detecção de fraudes em cartões de crédito (Quadro 1). Ryman-Tubb,

Krause e Garn (2018), em ampla revisão da literatura e *benchmark* da indústria, constataram que, apesar do elevado número de métodos publicados na área, apenas oito apresentavam desempenho suficiente para aplicação prática na indústria, evidenciando uma lacuna relevante entre os resultados reportados em artigos acadêmicos e a viabilidade de implantação em sistemas de produção.

Salekshahrezaee, Leevy e Khoshgoftaar (2023) avaliaram exclusivamente classificadores de *ensemble* (Random Forest, CatBoost, LightGBM e XGBoost), combinados a diferentes técnicas de extração de atributos e de amostragem, sobre uma base real de transações com cartão de crédito. Segundo os autores, a combinação entre a subamostragem aleatória e a extração de atributos via *autoencoder* convolucional produziu os melhores resultados entre as configurações testadas. Por não incluir modelos individuais na comparação, o estudo não permite afirmar a superioridade dos *ensembles* sobre outros algoritmos, mas documenta o comportamento desses classificadores diante do extremo desbalanceamento característico desse tipo de dados.

Em sentido diverso, Oliveira (2025), ao comparar Regressão Logística, Naive Bayes, Support Vector Machine e Árvore de Decisão na mesma base de transações europeias, obteve o melhor F1-Score (0,8247) com a Regressão Logística com limiar de decisão ajustado, à frente da Árvore de Decisão e do SVM. Em conjunto, esses resultados indicam que não existe um algoritmo universalmente superior, mas sim modelos mais ou menos adequados conforme o objetivo específico do sistema antifraude, seja priorizar a sensibilidade (*recall*), a precisão ou o equilíbrio entre ambas.

No contexto nacional, Souza e Bordin Jr. (2023) elaboraram um tutorial sobre a construção de modelos de reconhecimento de transações fraudulentas sob a perspectiva do aprendizado supervisionado, apresentando os conceitos dessa abordagem, as métricas de desempenho de classificação e as técnicas usuais de balanceamento de bases desbalanceadas.

### 4.2 Técnicas de balanceamento e pré-processamento de dados

O tratamento do desbalanceamento de classes constitui um dos temas mais recorrentes nos estudos analisados. Salekshahrezaee, Leevy e Khoshgoftaar (2023) compararam diferentes técnicas de amostragem, a subamostragem aleatória (RUS), o SMOTE e o SMOTE-Tomek, associadas a quatro classificadores de *ensemble* (Random Forest, CatBoost, LightGBM e XGBoost) e a técnicas de extração de atributos, a Análise de Componentes Principais (PCA) e o *autoencoder* convolucional (CAE). Segundo os autores, a subamostragem aleatória seguida do *autoencoder* convolucional produziu o melhor

desempenho, embora a ordenação das configurações varie conforme a métrica considerada (F1-Score ou AUC) e a escolha da técnica ideal dependa das características de cada base de dados.

A técnica SMOTE, originalmente proposta por Chawla et al. (2002), permanece como referência central nesse campo. Entre os estudos empíricos do corpus, contudo, a subamostragem aparece com destaque: Dal Pozzolo et al. (2018) treinam cada árvore de seu modelo em amostras balanceadas obtidas por subamostragem aleatória da classe majoritária, e o melhor resultado de Salekshahrezaee, Leevy e Khoshgoftaar (2023) foi obtido com subamostragem. Oliveira (2025), por sua vez, alcançou seu melhor resultado ajustando o limiar de decisão do classificador, estratégia que dispensa a alteração da base de treinamento.

Apesar de sua popularidade, o SMOTE pode gerar amostras sintéticas em regiões de sobreposição entre classes e introduzir ruído no processo de treinamento. Por esse motivo, variantes híbridas, como o SMOTE combinado a técnicas de limpeza de dados (SMOTE-Tomek), têm sido empregadas para melhorar a qualidade do conjunto de treinamento, removendo pares de exemplos considerados ruído ou casos de fronteira (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023).

Além do balanceamento propriamente dito, a forma como o problema é modelado mostra-se determinante para o desempenho dos modelos em condições reais. Dal Pozzolo et al. (2018) propõem uma modelagem realista da detecção de fraudes em cartões, que leva em conta o fluxo efetivo de operação dos sistemas antifraude, no qual apenas parte das transações é verificada por investigadores e os rótulos das demais chegam com atraso (*verification latency*). Os autores argumentam que o treinamento e a avaliação dos modelos devem refletir essas condições e a mudança do comportamento de fraudadores e clientes ao longo do tempo (*concept drift*), e não apenas características estáticas de cada transação isolada.

### 4.3 Métricas de avaliação de desempenho

A literatura analisada indica que a acurácia, métrica tradicionalmente utilizada em problemas de classificação, é inadequada para avaliar modelos de detecção de fraudes, dado o extremo desbalanceamento das classes. Na base de transações europeias, em que as fraudes representam 0,172% do total, um modelo que classificasse todas as transações como legítimas atingiria acurácia de 99,83% sem identificar uma única fraude; por isso, Oliveira (2025) destaca que a acurácia isolada é enganosa em dados desbalanceados.

Nesse sentido, métricas alternativas ganham centralidade nos estudos analisados, entre elas a precisão (proporção de transações classificadas como fraude que de fato o são), o *recall*

ou sensibilidade (proporção de fraudes reais corretamente identificadas) e o F1-Score, que representa a média harmônica entre precisão e *recall*, equilibrando ambos os objetivos. A área sob a curva ROC (AUC-ROC) e a área sob a curva de precisão-*recall* (AUPRC) também são utilizadas, sendo a segunda considerada mais informativa em contextos de forte desbalanceamento, por refletir de maneira mais fiel o desempenho do modelo sobre a classe minoritária (SAITO; REHMSMEIER, 2015). Dal Pozzolo et al. (2018), por sua vez, propõem avaliar os sistemas pela precisão dos alertas, isto é, a proporção de fraudes entre as transações de maior risco encaminhadas aos investigadores.

Hafez et al. (2025), em revisão sistemática de técnicas de inteligência artificial aplicadas à detecção de fraudes em cartões de crédito, reúnem estudos que reportam seus resultados por meio de diferentes métricas, como acurácia, precisão, sensibilidade e F1-Score, e apontam o desbalanceamento dos dados entre os principais problemas enfrentados pelos modelos. Diante dessa diversidade de indicadores, a escolha da métrica deve estar alinhada ao objetivo de negócio do sistema antifraude: em contextos nos quais o custo de investigar falsos positivos é elevado, prioriza-se a precisão; já em cenários nos quais deixar de detectar uma fraude representa perda financeira substancialmente maior, o *recall* assume papel central. Isso reforça a necessidade de que a avaliação considere não apenas métricas estatísticas, mas também o contexto operacional e econômico de aplicação.

### 4.4 Desafios, lacunas e tendências

A análise integrativa da literatura revela um cenário de avanços expressivos acompanhados de desafios ainda não plenamente resolvidos. Um primeiro desafio relevante refere-se à interpretabilidade dos modelos, especialmente daqueles baseados em *deep learning* e em métodos de *ensemble* complexos. Modelos de alta acurácia frequentemente funcionam como caixas-pretas, dificultando a compreensão dos motivos que levaram uma transação a ser classificada como fraudulenta, o que representa obstáculo tanto para a confiança de analistas humanos quanto para eventuais exigências regulatórias de explicabilidade (OLIVEIRA, 2025).

Um segundo desafio relaciona-se à escassez de dados reais e públicos para pesquisa. Dos três estudos empíricos do corpus, dois (SALEKSHAHREZAEE; LEEVY; KHOSHGOFTAAR, 2023; OLIVEIRA, 2025) utilizam a mesma base pública de transações europeias de setembro de 2013, o que levanta questionamentos sobre a generalização dos resultados para outros contextos geográficos, perfis de consumo e modalidades de fraude; o terceiro (DAL POZZOLO et al., 2018) depende de dados fornecidos por um parceiro

industrial. Nesse sentido, Ryman-Tubb, Krause e Garn (2018) defendem maior disponibilização de dados pela indústria. Adicionalmente, restrições de privacidade e sigilo bancário dificultam o compartilhamento de dados reais mais recentes e diversificados, o que também se relaciona às exigências legais de proteção de dados pessoais no tratamento de informações financeiras sigilosas (BRASIL, 2018).

A adaptação contínua a novos padrões de fraude constitui outro desafio central. Diferentemente de muitos problemas de classificação tradicionais, a detecção de fraudes lida com um adversário ativo, que ajusta suas estratégias à medida que os sistemas de defesa evoluem. Isso exige modelos capazes de aprendizado incremental ou de atualização periódica, sob risco de rápida obsolescência frente a novas modalidades de ataque (PENHA; MORAES; MELO, 2026; OLIVEIRA, 2025).

Por fim, observa-se uma desconexão relevante entre a sofisticação técnica alcançada em ambiente acadêmico e sua efetiva implementação em sistemas de produção (RYMAN-TUBB; KRAUSE; GARN, 2018), que precisam lidar com custo computacional, escalabilidade e monitoramento contínuo (OLIVEIRA, 2025). Estudos futuros devem, portanto, avançar na direção de modelos híbridos, que combinem a robustez estatística do *machine learning* tradicional com técnicas de *deep learning* e mecanismos de interpretabilidade, além de investigar arquiteturas capazes de operar de forma eficiente em cenários de processamento de dados em tempo real.

### 4.5 Decisão automatizada, explicabilidade e privacidade à luz da LGPD

Sistemas antifraude baseados em aprendizado de máquina tomam, em geral, decisões automatizadas que afetam diretamente o titular dos dados, como o bloqueio de um cartão ou a recusa de uma transação. No Brasil, essa operação está sujeita à Lei Geral de Proteção de Dados Pessoais (LGPD). O art. 20, na redação dada pela Lei nº 13.853/2019, assegura ao titular o direito de solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado de dados pessoais que afetem seus interesses, inclusive as destinadas a definir perfil de consumo e de crédito, bem como o direito de obter informações claras e adequadas sobre os critérios e os procedimentos utilizados, observados os segredos comercial e industrial. Quando essas informações não são fornecidas com base em tais segredos, a Autoridade Nacional de Proteção de Dados pode realizar auditoria para verificar aspectos discriminatórios no tratamento automatizado (BRASIL, 2018).

Esse dispositivo entra em tensão com a natureza de caixa-preta dos modelos mais acurados, discutida na seção anterior (OLIVEIRA, 2025). *Ensembles* complexos e redes

profundas dificultam informar ao titular, em linguagem compreensível, os critérios que levaram à classificação de uma transação fraudulenta. Essa dificuldade justifica o interesse por técnicas de explicabilidade aplicadas após o treinamento, como SHAP (LUNDBERG; LEE, 2017) e LIME (RIBEIRO; SINGH; GUESTRIN, 2016), e por modelos intrinsecamente mais interpretáveis, ainda que ao custo de eventual perda de desempenho. A ressalva do segredo comercial e industrial permite às instituições deixar de fornecer detalhes de seus modelos; nesse caso, porém, a lei prevê que a Autoridade Nacional de Proteção de Dados poderá realizar auditoria para verificar aspectos discriminatórios (art. 20, § 2º), e o direito do titular de solicitar a revisão da decisão permanece assegurado pelo caput do mesmo artigo.

No tocante à privacidade, os princípios da finalidade, da adequação, da necessidade, da segurança e da não discriminação, previstos no art. 6º da LGPD, orientam a construção desses sistemas: devem ser tratados apenas os dados necessários à finalidade de prevenção de fraudes, com medidas de segurança compatíveis e atenção para que os modelos não reproduzam vieses presentes nos dados históricos (BRASIL, 2018). Em favor das instituições, a lei admite o tratamento de dados para a proteção do crédito (art. 7º, X) e, no caso de dados pessoais sensíveis, para a garantia da prevenção à fraude e da segurança do titular nos processos de identificação e autenticação de cadastro em sistemas eletrônicos (art. 11, II, "g"), resguardados os direitos do titular (BRASIL, 2018).

Do ponto de vista da pesquisa, a revisão de Hafez et al. (2025) inclui estudos que buscam aliar desempenho e proteção da privacidade e apontam o aprendizado federado, que permite treinar modelos de *deep learning* em fontes de dados distribuídas sem comprometer a privacidade, como direção futura. Nos demais estudos analisados nesta revisão, contudo, a ênfase recai sobre métricas de desempenho preditivo, sendo menos frequente a discussão explícita de requisitos legais de explicabilidade e privacidade. Essa lacuna indica a necessidade de que futuras pesquisas incorporem, desde a concepção dos modelos, critérios de conformidade com a LGPD, e não apenas indicadores estatísticos de desempenho.

## 5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo analisar, por meio de revisão bibliográfica, as principais técnicas de *machine learning* aplicadas à detecção de fraudes em transações financeiras. A revisão da literatura evidenciou que algoritmos de *ensemble*, como Random Forest e *boosting*, são recorrentes nos trabalhos analisados, mas que modelos mais simples podem apresentar desempenho competitivo, como mostra o resultado obtido com a Regressão

Logística em um dos estudos comparativos do corpus, o que reforça a inexistência de um algoritmo universalmente superior. Técnicas de *deep learning*, embora demandem maior volume de dados e capacidade computacional, vêm ganhando espaço por sua habilidade de capturar padrões temporais e relações complexas nas transações.

O desbalanceamento extremo entre transações legítimas e fraudulentas foi identificado como um dos principais desafios metodológicos enfrentados pelos pesquisadores da área, sendo tratado por meio da subamostragem, do SMOTE e de suas variantes híbridas e do ajuste do limiar de decisão. Constatou-se ainda que a avaliação de desempenho desses modelos exige métricas específicas, como F1-Score, AUC-ROC, AUPRC e a precisão dos alertas, uma vez que a acurácia isolada se mostra insuficiente e potencialmente enganosa nesse contexto.

Entre as principais lacunas identificadas, destacam-se a limitada interpretabilidade dos modelos mais sofisticados, a escassez de conjuntos de dados reais, públicos e diversificados para validação, e a dificuldade de manter os sistemas atualizados diante da constante evolução das estratégias fraudulentas. Observou-se, ainda, que o direito à revisão de decisões automatizadas e as exigências de privacidade previstos na LGPD são pouco explorados nos estudos de desempenho, o que reforça a importância de modelos mais explicáveis. Adicionalmente, verificou-se uma desconexão relevante entre os avanços acadêmicos e sua efetiva aplicação em ambientes de produção, sujeitos a restrições operacionais de tempo real e escalabilidade que nem sempre são consideradas nos estudos experimentais analisados.

Como limitação deste estudo, reconhece-se que a revisão bibliográfica realizada não seguiu um protocolo sistemático rígido e se apoiou em um corpus de oito trabalhos, o que pode ter restringido a abrangência do levantamento em comparação a revisões sistemáticas formais. Recomenda-se, para trabalhos futuros, a condução de revisões sistemáticas com protocolos como o PRISMA (PAGE et al., 2021), bem como estudos empíricos que apliquem e comparam, de forma prática, as técnicas aqui discutidas sobre bases de dados nacionais, contribuindo para a validação dessas abordagens no contexto brasileiro. Em síntese, o aprendizado de máquina oferece contribuição relevante à detecção de fraudes, mas sua efetividade prática depende de dados representativos, de métricas alinhadas ao negócio, de modelos explicáveis e de atualização frequente diante do caráter adversarial do problema.

## REFERÊNCIAS

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, 15 ago. 2018. Redação consolidada com as alterações da Lei nº 13.853, de 8 de julho de 2019. Disponível em:

https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 21 out. 2026.

CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. Philip. SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321-357, 2002. DOI: 10.1613/jair.953.

DAL POZZOLO, Andrea; BORACCHI, Giacomo; CAELEN, Olivier; ALIPPI, Cesare; BONTEMPI, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy. **IEEE Transactions on Neural Networks and Learning Systems**, v. 29, n. 8, p. 3784-3797, 2018. DOI: 10.1109/TNNLS.2017.2736643.

HAFEZ, Ibrahim Y.; HAFEZ, Ahmed Y.; SALEH, Ahmed; ABD EL-MAGEED, Amr A.; ABOHANY, Amr A. A systematic review of AI-enhanced techniques in credit card fraud detection. **Journal of Big Data**, v. 12, art. 6, 2025. DOI: 10.1186/s40537-024-01048-8.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: GUYON, I. et al. (ed.). **Advances in Neural Information Processing Systems 30**. Red Hook: Curran Associates, 2017.

OLIVEIRA, Vinicius Dias. Métodos de machine learning na detecção de fraude em cartão de crédito: um estudo comparado. **Revista Científica ACERTTE**, v. 5, n. 9, e59265, 2025. DOI: 10.63026/acertte.v5i9.265.

PAGE, Matthew J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, v. 372, n71, 2021. DOI: 10.1136/bmj.n71.

PENHA, Yann Benedito Araújo; MORAES, Nathaniel Luiz Cordeiro; MELO, Dadilton Bastos. Uso de aprendizado de máquina na detecção de fraudes financeiras. **Boletim de Conjuntura (BOCA)**, v. 25, n. 78, e8251, 2026. DOI: 10.56238/bocav25n78-020.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why should I trust you?": explaining the predictions of any classifier. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 22., 2016, San Francisco. **Proceedings** [...]. New York: ACM, 2016. p. 1135-1144. DOI: 10.1145/2939672.2939778.

RYMAN-TUBB, Nick F.; KRAUSE, Paul; GARN, Wolfgang. How artificial intelligence and machine learning research impacts payment card fraud detection: a survey and industry benchmark. **Engineering Applications of Artificial Intelligence**, v. 76, p. 130-157, 2018. DOI: 10.1016/j.engappai.2018.07.008.

SAITO, Takaya; REHMSMEIER, Marc. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **PLoS ONE**, v. 10, n. 3, e0118432, 2015. DOI: 10.1371/journal.pone.0118432.

SALEKSHAHREZAEE, Zahra; LEEVY, Joffrey L.; KHOSHGOFTAAR, Taghi M. The effect of feature extraction and data sampling on credit card fraud detection. **Journal of Big Data**, v. 10, art. 6, 2023. DOI: 10.1186/s40537-023-00684-w.

SOUZA, Daniel Henrique Miguel de; BORDIN JR., Claudio J. Detecção de fraude de cartão de crédito por meio de algoritmos de aprendizado de máquina. **Revista Brasileira de Computação Aplicada**, v. 15, n. 1, p. 1-11, 2023. DOI: 10.5335/rbca.v15i1.13790.

TOSTA, Pedro Lucas Maranini; DIAS, Jonatas Cerqueira. Detecção de fraudes em transações bancárias utilizando inteligência artificial. **Revista Processando o Saber**, v. 17, n. 1, p. 21-37, 2025. DOI: 10.5281/zenodo.15477217.