UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE COMPUTAÇÃO

Exame de Qualificação de Mestrado

31 de Janeiro de 2024

Identificação de anomalias em transações financeiras e avaliação de ações preventivas no ecossistema de pagamentos

Candidato: Marcos Vinícius Piaia

Orientador: Prof. Dr. Zanoni Dias

Coorientador: Prof. Dr. Hélio Pedrini

Sumário

Resumo

O sistema financeiro atual, especialmente em nossa realidade brasileira, possui um grau de flexibilidade, dinamismo e complexidade muito grandes. Neste cenário, transações que antes aconteciam em minutos, como a TED [9], e ainda possuíam um certo grau de interação com um revisor humano, estão, agora, sujeitas a agentes autônomos que buscam atender à conveniência e velocidade exigidas por estas novas formas de transações financeiras, como o Pix [8]. Além disso, mesmo transações tradicionais como operações via cartão de crédito possuem um ecossistema cada vez mais complexo, no qual não apenas bancos e grandes operadores de crédito estão envolvidos, mas também novas organizações privadas que passaram a operar dentro do sistema financeiro. Infelizmente, o desenvolvimento de tais facilidades colaboraram não apenas para o dinamismo e flexibilidade, mas também para o aumento dos casos de ações fraudulentas. Com tais transações dentro deste ecossistema, vê-se que existe um grande desafio com relação à velocidade com a qual se identifica uma ação indevida e também na rapidez com que medidas são tomadas com os diversos participantes do ecossistema, buscando minimizar a ação e o resultado final da fraude. De posse deste objetivo, propõe-se aqui um fluxo de operações valendo-se de modelos de aprendizado não supervisionado para garantir a rápida identificação de transações suspeitas, automatizando o tratamento das mesmas com abordagens envolvendo bloqueios automáticos e ações que miram o combate às tentativas de fraude envolvendo ações de engenharia social, buscando combater elementos como a personificação de usuários lícitos por fraudadores atacando o sistema.

Capítulo 1

Introdução

Este capítulo caracteriza o problema a ser investigado, apresenta os principais objetivos e contribuições do trabalho, apresenta suas questões de pesquisa, bem como a organização do texto.

1.1 Caracterização do Problema

O sistema financeiro atual, especialmente na nossa realidade brasileira, possui grande flexibilidade e complexidade. Tais características são devido a fatores como o alto grau de regulamentação desta área e também aos diversos elementos que podem compô-la. Esta última sendo o fator essencial de promoção de sua flexibilidade.

Neste sistema, um usuário é capaz de realizar uma transação para outro participante em segundos, tempo este que é regulamentado pela entidade que monitora o sistema financeiro brasileiro (BACEN) [5], gerando punições para as entidades que não respeitam esses requisitos [7]. Ao mesmo tempo, um usuário pode realizar um pagamento via um cartão de crédito atrelado a uma instituição de pagamento que oferece uma conta digital garantida por um banco tradicional, mas cujas transações do cartão são efetivadas por uma Sociedade de Crédito Direto (SCD) [6], o que nos permite ver a quantidade de participantes envolvidos em uma transação, a pluralidade de casos de uso atendidos e os desafios que transitam entre o suporte ao tempo mínimo de transações e a orquestração apropriada entre os vários participantes da mesma. Note que tanto no cenário do Pix, quando na transação via cartão, uma fraude pode apresentar consequências catastróficas para todos os envolvidos, visto que um usuário vítima desta fraude pode ter sua conta e seu crédito comprometidos em segundos e que as instituições envolvidas podem sofrer grandes prejuízos até que a fraude seja detectada.

Dessa forma, vemos que o desafio encontra-se não só na velocidade com a qual se identifica uma ação indevida no sistema, como também na rapidez com que medidas sejam tomadas com os diversos participantes deste para coibi-la. Pesquisando sobre trabalhos relacionados, identificamos vários tratando da caracterização da transação fraudulenta valendo-se de estratégias supervisionadas [17], semi [33] e não supervisionadas [31], buscando a criação de processos e modelos para realizar esta identificação.

Nosso trabalho dedica-se a uma revisão destes para caracterizar sua eficácia com variações

dos dados utilizados para validações e testes com o uso de outras bases de transações, bem como com adaptações nos modelos para melhoria de seu desempenho. Após este trabalho, nosso objetivo é o desenvolvimento de um fluxo de operações, valendo-se do estudo realizado, com o objetivo de garantir a rápida identificação de transações suspeitas e a automatização do tratamento destas com abordagens envolvendo bloqueio automático de transações, culminando com o acompanhamento e medição dos impactos destas ações no histórico de tentativas de fraude do sistema como um todo.

1.2 Objetivos e Contribuições

Para desenvolver a metodologia proposta, alguns objetivos específicos devem ser alcançados:

- Levantamento bibliográfico e estudo das abordagens utilizadas para a identificação de transações financeiras suspeitas;
- Estudo comparativo entre bases de dados encontradas para o estudo;
- Realização de experimentos;
- Avaliação e comparação do método proposto com outras abordagens disponíveis;
- Documentação e publicação dos resultados.

Este projeto visa contribuir com a definição e o desenvolvimento de uma metodologia para a identificação de fraudes em transações financeiras, tratando-as como anomalias no conjunto de transações lícitas, tal que este processo de identificação possa ser aplicado ao ecossistema de pagamentos, fortalecendo a segurança do usuário e provendo uma estrutura de fácil aplicação ao gestor financeiro.

1.3 Hipóteses do Trabalho

As seguintes hipóteses serão investigadas neste trabalho:

- Na busca por evitar a criação de dados sintéticos, a manipulação dos dados de entrada para facilitar a identificação do comportamento característico da normalidade nas transações poderá viabilizar e melhorar o resultado em algoritmos que dependem de bases mais balanceadas.
- Um dos modelos propostos apresentará uma capacidade de generalização eficaz nas diversas bases consideradas para o estudo.
- Esse modelo será capaz de generalizar a identificação de transações fraudulentas em diferentes bases de dados sem a necessidade da criação de dados sintéticos para o balanceamento da base.

1.4 Organização do Texto

O Capítulo 2 descreve os conceitos e técnicas relevantes relacionados ao tema sob investigação. O Capítulo 3 descreve a metodologia proposta, a base de dados, as métricas de avaliação e os recursos computacionais que serão empregados no desenvolvimento do projeto. O Capítulo 4 apresenta alguns resultados preliminares. O Capítulo 5 apresenta o plano de trabalho e o cronograma de execução das atividades.

Capítulo 2

Revisão Bibliográfica

O uso de técnicas envolvendo inteligência artificial e aprendizado de máquina para a identificação de fraudes é um assunto amplamente estudado, no qual, o principal interesse é a natureza do evento que a caracteriza, ou seja, a anomalia. Na busca por identificar e entender esse evento, no trabalho de Pang et al. [31] encontramos um estudo buscando caracterizar os diversos aspectos do problema e os pontos-chave a serem analisados para a busca de uma solução. Seus autores promovem uma caracterização e classificação dos diversos métodos construídos com o uso de técnicas de aprendizado profundo, buscando avaliar as principais características de cada um deles e seus resultados.

Nosso estudo parte da premissa que uma fraude é uma anomalia na base de dados de transações e ainda que considere as informações levantadas por Pang et al. [31], busca complementar algumas das abordagens indicadas focando na análise do comportamento de estratégias selecionadas de identificação de anomalias no estudo com dados de transações, validando premissas sobre a efetividade destes modelos quando aplicados em alguns conjuntos de dados públicos [18, 20, 22, 23].

2.1 Identificação de Anomalias

Podemos considerar os seguintes pontos como elementos que compõem a complexidade da busca pela identificação de anomalias [31]:

- Em geral o evento que caracteriza a anomalia se mantém desconhecido até o seu acontecimento;
- Essas anomalias podem se apresentar agrupadas em classes bastante heterogêneas, dificultando sua generalização;
- Considerando-se a existência de tipos muito diversos de anomalias, ressaltam-se os seguintes:
  - Anomalias pontuais, onde instâncias individuais dos dados são consideradas anômalas com respeito ao restante da base.
  - Anomalias condicionais, nas quais ainda se consideram instâncias individuais dos dados como sendo anômalos não com relação a uma característica de toda base,

mas sim, como consequência gerada a partir de um contexto, caracterizado por uma combinação de outros eventos da mesma base.

  - Anomalias de grupo, nos quais se tratam grupos de dados como anômalos com relação a outros da mesma base. Por exemplo, um conjunto de contas falsas em uma plataforma de mídia social. A análise individual destas contas pode mostrar perfis válidos com ações válidas, enquanto estas, agrupadas, revelam uma ação fraudulenta.

- Anomalias são eventos naturalmente raros, criando bases de dados desbalanceadas;

Considere o cenário no qual uma transação de valor expressivo pode não ser considerada fraudulenta, pois ela se alinha com o perfil de transações na plataforma, no entanto, a mesma transação, quando analisada juntamente com uma série de outras em sequência, em intervalos muito próximos, pode sim caracterizar uma ação decorrente de um atacante na conta, o mesmo valendo para uma transação individual, mas realizada em uma localidade e/ou horários muito discrepantes da média para o mesmo usuário ou base de dados. O mesmo se aplica quando analisamos o comportamento de usuários de uma plataforma financeira, que, individualmente, estão realizando transações lícitas em termos de valores, no entanto, sua movimentação, quando analisada em conjunto, mostra uma exploração do fluxo financeiro para processos de lavagem de dinheiro entre várias contas.

De posse destes elementos sobre a complexidade do ambiente, a qual consideramos nosso cenário extremamente aderente em função dos exemplos citados, um conjunto de desafios na identificação de anomalias foi levantado por Pang et al. [31]:

- Baixa taxa de recuperação de detecção de anomalias, ou seja, os métodos atuais ainda apresentam uma alta taxa de falsos positivos quando aplicados a bases reais.
- Detecção de anomalias em bases de dados de grande dimensionalidade e/ou dados não-independentes, ou seja, aquelas com elementos que apresentam discrepâncias que tendem a ser óbvias quando considerando apenas um pequeno conjunto de características ou em listagens filtradas da informação, mas que se perdem quando avaliados com todo o conteúdo disponível.
- Aprendizado eficiente da normalidade/anormalidade dos dados, que é um desafio no aprendizado supervisionado devido à escassez de dados classificados, e também no aprendizado não supervisionado devido ao desbalanceamento das bases, onde existem, no geral, poucos grupos de elementos que um revisor poderia considerar como anômalo. Nesse contexto, ganham força estratégias como o aprendizado semi-supervisionado, que se vale de uma pequena gama de dados classificados para realizar o aprendizado da rede, ou ainda, o aprendizado fracamente supervisionado [46], contemplando cenários com ainda menos informações propriamente classificadas ou até mesmo erroneamente classificadas para treinamento e classificação.
- Resiliência ao ruído dos dados, ou seja, dados mal classificados quando tratamos de aprendizado supervisionado e bases com grupos pouco representativos de informação quando falamos dos casos de aprendizados semi, fracamente e não supervisionados.

- Detecção de anomalias complexas, na qual a identificação passa pela análise de dados em diversos formatos e de grande complexidade.
- Explicabilidade de uma identificação, ou seja, além de realizar o feito de encontrar a anomalia, ter um ecossistema capaz de prover insumos que permitam a um analista externo identificar o caminho seguido pelo sistema para a identificação de uma determinada instância ou grupo de dados como anômalos.

De acordo com Pang et al. [31], com a identificação dos desafios propostos anteriormente, vemos a definição do problema de identificação de anomalia com aprendizado profundo como um conjunto $X = \{x_1, x_2, ..., x_N\}$ de eventos, onde $x_i \in \mathbb{R}^D$ é representado por um grupo de elementos que definem as características daquele evento, no qual, $Z \in \mathbb{R}^k (k \ll N)$ é a representação do espaço de anomalia a ser estudada. Dessa forma, o objetivo é encontrar $\phi(\cdot) : X \to Z$ ou $\tau(\cdot) : X \to \mathbb{R}$, na qual ϕ é uma função capaz de identificar o conjunto de características que identificam uma anomalia e posteriormente podem ser aplicadas a um classificador e τ, por outro lado, já representa uma função que fornece diretamente um valor que caracteriza uma pontuação para a identificação do evento anômalo em X.

Nesta abordagem, ϕ e τ são funções de mapeamento em uma rede com $H \in \mathbb{N}$ camadas escondidas com suas respectivas matrizes de peso $\Theta = \{M^1, M^2, ..., M^H\}$. Tal rede pode possuir diversas arquiteturas envolvendo camadas totalmente conectadas, de agrupamento, convolucionais, recorrentes, etc., e utilizando as mais diversas funções de ativação, como ReLU (*Rectified Linear Unit*) e outras alternativas previamente definidas e exploradas [41]. Com isso, Pang et al. [31] propõem uma estrutura de classificação destes diversos métodos, subdivindo-os nas seguintes classes:

- Aprendizado profundo para extração de características.
- Aprendizado de características de normalidade.
- Aprendizado de características de anormalidade.
- Aprendizado fim-a-fim da pontuação de anomalias.

A seguir exploramos alguns pontos, indicando o que acreditamos ser relevante para análise de nosso problema e de importante contribuição para a extensão do conhecimento agregado pela análise realizada em Pang et al. [31] com foco no cenário de anomalias do ambiente financeiro.

2.1.1 Aprendizado Profundo para Extração de Características

Os métodos agrupados aqui lançam mão do aprendizado profundo para a extração de poucas dimensões de características a partir de bases de dados com pouca linearidade e/ou alta dimensionalidade, separando os processos de extração de características da pontuação para classificação da anomalia. Com isso, é comum a reutilização de redes pré-treinadas para a identificação de características relevantes e, na sequência, a aplicação de classificadores como:

- Um único modelo de SVM (*Support Vector Machines*) aplicado diretamente para definir o evento como anomalia [2].

- Um conjunto de vários modelos de SVM para a classificação de vários tipos de anomalias, buscando não apenas destacar o evento, mas classificá-lo quanto ao seu tipo. Uma abordagem como esta pode ser interessante quando os dados anômalos possuem características tão diferentes que podem se perder no conjunto de dados, um cenário que este grupo de planos de corte pode ajudar a identificar adequadamente.

Um exemplo com dados tabulares foi demonstrado com a utilização do conceito denominado *Deep Belief Networks* ou Redes de Crenças Profundas [11], utilizado para a redução de dimensionalidade dos dados, para então culminar com o treinamento de um ou vários modelos SVM para a identificação e classificação de anomalias.

2.1.2 Aprendizado de Características de Normalidade

Nos modelos agrupados nesta classificação, o processo de identificação de características e geração da pontuação classificatória da anomalia acontecem juntos:

$$\Theta^*, W^* = argmin_{\Theta,W} \sum_{x \in X} \ell(\psi(\phi(x, \Theta), W)) \tag{2.1}$$

onde ϕ é a função de otimização que leva x a Z, ψ é a função custo dependente de W que reforça o aprendizado das características de Z e ℓ é a função custo utilizada para atingir os melhores modelos de ψ e ϕ que deverão ser usados para compor a pontuação classificatória de um evento no modelo:

$$S_x = f(x, \phi_\Theta, \psi_W) \tag{2.2}$$

Considerando as arquiteturas de rede utilizadas nestes métodos podemos agrupa-los da seguinte forma:

- Auto-Codificadores.
- Redes Generativas Adversariais.
- Modelagem de Previsibilidade.
- Classificação Autossupervisionada.

2.1.2.1 Auto-Codificadores

Auto-codificadores (*autoencoders*) são modelos com foco na redução de dimensionalidade com a preservação de conteúdo para a reconstrução da informação. A premissa é que um dado considerado normal pode ser reconstruído com uma taxa muito baixa de erro, o que não se repete com dados anômalos, de forma que a pontuação classificatória de um evento pode ser calculada a partir do próprio erro da função de custo para definir se o mesmo é ou não uma anomalia.

A aplicação destas técnicas em dados tabulares foi validada por Hawkins et al. [16] com uma estrutura que foi denominada Rede Neural de Replicadores (RNN) e posteriormente testada com séries temporais por Kieu et al. [24].

A RNN trata-se de uma rede neural de múltiplas camadas com neurônios totalmente conectados e considerando três destas camadas escondidas entre outras duas. Uma destas duas é a de entrada, na qual cada neurônio representa uma característica do dado a ser avaliado e a outra a camada de saída, representando a reconstrução do vetor de características do evento avaliado. A Figura 2.1 ilustra o esquema de tal rede.

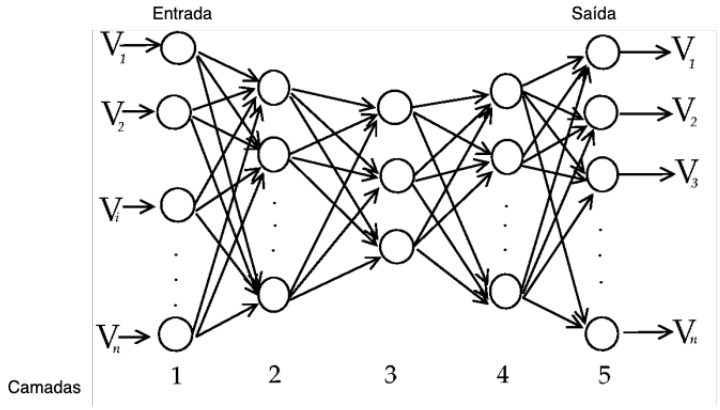


Figura 2.1: Modelo reproduzido a partir do proposto por Hawkins et al. [16] para uma RNN.

Vale ressaltar que além da arquitetura mais simples da rede, a definição da quantidade de neurônios nas camadas de entrada, saída e naquelas escondidas são resultados de estudos empíricos nos conjuntos de dados avaliados, não existindo uma regra geral para a definição destes valores.

Auto-codificadores também foram usados por Zhou et al. [45] com RPCA [3] para a redução de ruído na decodificação dos dados, apresentando resultados interessantes na prevenção da falsa identificação de anomalias.

2.1.2.2 Redes Generativas Adversariais

Redes Generativas Adversariais (*Generative Adversarial Networks* ou GANs) consideram aprendizado de uma característica do espaço latente gerado por uma rede G de forma que tal espaço representa a normalidade. Dessa forma, o objetivo é gerar um elemento $z = G(x)$ e que a medida de similaridade entre este elemento e o item em análise represente a pontuação classificatória da anomalia. Como o espaço G deve acompanhar a mesma distribuição do dado considerado normal, a seguinte função objetivo é aplicada:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_x}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{2.3}$$

onde G e D são redes denominadas gerador e discriminador, parametrizadas por Θ_G e Θ_D e V é a função valor. Com isso, para encontrar o melhor z, cada x é aplicado em duas funções custo:

- Custo residual do gerador:

$$\ell_R(x, z_\gamma) = \|x - G(z)\| \tag{2.4}$$

- Custo residual do discriminador:

$$\ell_R(x, z_\gamma) = \|h(x) - h((G(z))\| \tag{2.5}$$

onde γ é o índice utilizado na iteração para a busca do melhor z, e h é o mapeamento de característica da camada intermediária do discriminador.

A busca começa com a escolha aleatória de z e a atualização de seu valor pela composição da função custo dada por $(1-\alpha)\ell_R(x, z_\gamma) + \alpha\ell_{fm}(x, z_\gamma)$, com α sendo um hiperparâmetro. Por fim, com a identificação de z, a pontuação classificatória para x surge a partir da medida de similaridade entre x e z a partir da equação:

$$S_x = (1-\alpha)\ell_R(x, z^*_\gamma) + \alpha\ell_{fm}(x, z^*_\gamma) \tag{2.6}$$

No entanto, este método apresenta-se computacionalmente muito custoso quando considera-se o processo de descoberta de z. Para resolver isso, outras iniciativas foram realizadas destacando-se entre elas a EBGAN [42], a qual, baseando-se na abordagem da BiGAN [10] e em um método chamado ALAD [43], adicionou dois outros discriminadores que buscam encontrar os pares (x, x) de $(x, G(E(x)))$ e (z, z) de $(z, G(E(z)))$.

Outra abordagem, buscando também melhorar a performance com a utilização da GAN na detecção de anomalias pode ser encontrada em Akcay et al. [1], onde foram utilizados codificadores e decodificadores para melhorar o mapeamento dos valores de x no espaço latente criado a partir de G. Dentre as desvantagens de utilização destas redes, encontra-se a possibilidade de não convergência e colapso do modelo, como citado por Metz et al. [28], dificultando o treinamento em grandes bases de dados.

2.1.2.3 Modelagem de Previsibilidade

Modelos baseados em métodos de previsibilidade aprendem a representação da característica de uma instância atual dos dados considerando uma janela temporal anterior de elementos como contexto. Dessa forma, considera-se que dados normais são aderentes à representação em uma sequência, de forma que seriam previsíveis, ao contrário dos dados anômalos, os quais seriam imprevisíveis. Logo, o erro identificado na previsão seria a pontuação classificatória de identificação da anomalia. Identificamos usos desta análise de sequência em traduções [37] e dados de tráfego [26], ambos com boa eficácia. Especialmente, estes modelos são interessantes quando consideramos sua forma tabular com extensas quantidades de dados temporais, características que também são encontradas nos conjuntos de transações financeiras que temos.

2.1.2.4 Classificação Autossupervisionada

Classificação autossupervisionada baseia-se na utilização de múltiplos classificadores para o aprendizado de características específicas de uma base dados e na subsequente composição dos valores de erro destas com os valores reais durante o treinamento. Para a detecção da anomalia, estes valores seriam a pontuação classificatória do item em identificação. O modelo baseia-se em dois pontos descobertos durante a avaliação da base de dados:

- A magnitude do gradiente gerado por classes nos dados considerados normais é substancialmente maior que nos dados anômalos;
- A direção da atualização da rede também sofre com um desvio em direção à classificação dos dados normalizados.

A devantagem deste método é que ele é altamente dependente de transformações aplicadas sobre os dados para o próprio aprendizado de características, sendo portanto, mais adequado para o tratamento de imagens, onde tais transformações podem ser aplicadas sem que o contexto da informação na imagem seja alterada.

2.1.3 Aprendizado de Características de Anormalidade

Até o momento, os métodos previamente mencionados consideravam a caracterização de informações buscando identificar a normalidade, já os métodos a seguir buscam o aprendizado de características que identifiquem a anormalidade. De forma geral, os métodos a seguir podem ser traduzidos como:

$$(\Theta^*, W^*) = argmin_{\theta,w} \sum_{x \in X} l(f(\phi(x, \theta), w)) \tag{2.7}$$

$$S_x = f(\phi(x, \theta^*), w^*) \tag{2.8}$$

onde f é aplicada no espaço de representação dos dados e no cálculo da pontuação classificatória da anomalia, sendo esta uma importante característica destes métodos, visto que nos anteriores a medida utilizada na criação desta pontuação era aplicada apenas após a heurística de aprendizado da rede. A expectativa destes métodos é que a incorporação da função f otimize sua performance no cálculo da pontuação classificatória de anomalias.

Nestes conjuntos de métodos, identificamos três subgrupos baseados na métrica de avaliação de características, que serão indicados a seguir.

2.1.3.1 Medições Baseadas em Distância

Utilizando métodos baseados em distâncias entre classes, as abordagens de identificação de anomalias neste grupo buscam aproveitar a extensa literatura existente, a capacidade de aprendizado com dados de baixa dimensionalidade e a adaptabilidade para a busca de padrões de anomalias já conhecidas dos dados revisados com a criação de processos robustos de identificação de tais informações anômalas.

Problemas como a quantidade de dados anômalos para treinamento são enfrentados através de abordagens como a identificação de anomalias baseado na distância aleatória (*random distance-based outlier detection*), a qual define a pontuação de classificação de um objeto como a distância deste em relação a um subconjunto aleatório do grupo de elementos anômalos [30] ou mesmo o desenvolvimento de redes que não utilizam nenhum dado previamente classificado no aprendizado e fazem uso da estimativa destas distâncias para então gerar a pontuação classificatória do objeto [40].

2.1.3.2 Medições para Classificação com Classe Única

Estes métodos utilizam redes neurais para criarem projeções das bases de dados sendo avaliadas com menor dimensionalidade e a partir destas, aplicarem abordagens como SVM de classe única [36] e SVDD [38] para identificar anomalias.

A premissa desses casos é que a base de dados pode ser englobada em uma única grande classe a qual o dado anômalo não tem participação. A principal vantagem é a extensa gama de estudos sobre os métodos envolvendo SVM, dos quais podemos citar a implementação feita por Ruff et al. [35], na qual, o problema da pouca quantidade de dados classificados foi tratado através de uma abordagem semi-supervisionada com a expansão do conjunto existente de dados classificados como anômalos utilizando processos de regularização que buscavam expandir o conjunto, mas manter as propriedades estatísticas de seus elementos.

A desvantagem neste caso é a baixa performance em bases onde aqueles dados considerados normais apresentam uma distribuição mais complexa, tornando difícil a caracterização desta grande classe que serviria como base para a construção da anormalidade.

2.1.3.3 Medições Baseadas em Agrupamentos

Os métodos deste grupo buscam aprender representações dos dados nas quais as anomalias possuem desvios bastante claros dos agrupamentos identificados por estas novas representações do espaço de características, dessa forma, o Agrupamento Profundo (*Deep Clustering*), tem por objetivo o aprendizado sob medida de um determinado agrupamento que será a base de classificação da anomalia, de forma que o algoritmo utilizado para tal é a parte crucial destes métodos.

No geral, consideram-se aqui duas importantes premissas: (i) uma representação adequada do espaço de características leva a um bom agrupamento, o que, por sua vez, leva a um bom aprendizado deste espaço; e (ii) representações que são otimizadas para um determinado algoritmo de agrupamento, não necessariamente são apropriadas para outros algoritmos.

Uma arquitetura de rede destinada ao tratamento deste tipo de método considera dois módulos principais: um que realiza o agrupamento na passagem de propagação dos pesos pela rede e outro que realiza o aprendizado do espaço de representações utilizando os dados de agrupamento como pseudo-classes daqueles na retro-propagação.

De forma geral, esta abordagem pode ser representada como:

$$\alpha\iota_{clu}(f(\phi(x,\Theta);W),y_x)+\beta\iota_{aux}(\chi) \tag{2.9}$$

onde ι_{clu} é uma função de custo do agrupamento, na qual ϕ é a função de aprendizado parametrizada por Θ, f é a função de atribuição da classe do agrupamento parametrizada por W e y_x representa a pseudo-classe atribuída para pelo processo de agrupamento; ι_{aux} é uma função de custo não atrelada ao agrupamento e considerada para a aplicação de regras adicionais no aprendizado do espaço de representações; e α e β são dois hiperparâmetros destinados a estabelecer a importância dada às duas funções de custo consideradas.

Um importante elemento destes métodos é a base de dados utilizada para o seu treinamento, visto que o aprendizado do espaço de características está intimamente ligado à especi-

alização do algoritmo de agrupamento, de forma que dados contaminados com representações de anomalias no treinamento pode levar a uma classificação mais pobre deste espaço de normalidade e comprometer a classificação da anomalia. A utilização de uma segunda função de custo busca minimizar essas contaminações e, assim, aumentar a abrangência da utilização destes métodos.

Com isso, a vantagem que pode ser verificada aqui seria a extensa literatura disponível com a criação de diversas estruturas de agrupamento, o que facilita a busca pelo correto tratamento dos dados, ao mesmo tempo que a sua grande desvantagem é a necessidade de uma revisão adequada das bases de treinamento para minimizar sua contaminação com dados anômalos que podem comprometer a classificação adequada do espaço de características, mesmo considerando as funções de custo exploradas com $\beta\iota_{aux}(\chi)$.

2.1.4 Aprendizado Fim-A-Fim da Pontuação de Anomalias

Neste conjunto de métodos, a principal característica identificada é o foco na obtenção direta de uma classificação de anomalia já na saída do modelo de rede considerado, focando no tratamento da função de custo como forma de criação do escalar que irá representar a pontuação classificatória da anomalia. Ainda que existam algumas similaridades com os métodos considerados nas Seções 2.1.2 e 2.1.3, nestes a diferença fundamental encontra-se no fato que não existe um esforço prévio para o aprendizado do espaço de características que será base para a classificação da anomalia, o que torna-se uma vantagem quando considerado o tempo investido no treinamento destas redes. Formalmente, pode-se definir estes métodos como $\tau(\cdot;\Theta) : \chi \to \mathbb{R}$ e a representação geral do processo através das equações:

$$\Theta^* = argmin_\Theta \sum_{x\in\chi} \iota(\tau(x;\Theta)) \tag{2.10}$$

$$S_x = \tau(x;\Theta^*) \tag{2.11}$$

Além disso, é possível agrupá-los nas seguintes categorias: modelos de ranqueamento, modelos baseados em distribuições conhecidas dos dados e modelos de máxima verossimilhança.

2.1.4.1 Modelos de Ranqueamento

Este grupo de métodos busca aprender um modelo de ranqueamento tal que seja possível ordenar as instâncias de dados com base em um valor ordinal observável com relação direta ou relativa ao seu grau de anormalidade na base de dados considerada. A premissa aqui é a possibilidade ordenar os dados avaliados de acordo com uma gradação de sua anormalidade.

Nestes métodos, uma linha de pesquisa utiliza a combinação de dois elementos do conjunto e usa este relacionamento para definir a gradação de anormalidade de um relacionamento de dados. Esta abordagem foi aplicada com uma solução totalmente não supervisionada para o tratamento dos dados e identificação das anomalias [34], como também considerando a utilização de um pequeno conjunto de dados classificados para o treinamento do modelo que também considerava esta avaliação com pares de dados combinados [32].

As principais vantagens destes métodos seriam que as pontuações classificatórias de anormalidade poderiam ser otimizados diretamente a partir da função de custo e que eles, em geral, são livres da necessidade de uma classificação de anomalia por imporem a existência de uma relação necessariamente fraca entre duas informações, o que seria característico apenas de pares de informações compostos por um dado normal e um dado anômalo. As desvantagens ficam por conta da baixa eficácia identificada com métodos totalmente não supervisionados [34] em comparação com o modelo semi-supervisionado [32] e em específico com a natureza do problema que estamos avaliando em nosso cenário, o qual trata de dados tabulados de baixa dimensionalidade, realidade bem diferente daqueles vistos nos estudos mencionados acima com foco em processamento de vídeos.

2.1.4.2 Modelos Baseados em Distribuições Conhecidas dos Dados

Estes modelos consideram a abordagem, na qual uma distribuição conhecida dos dados é aplicada para codificar e guiar o aprendizado da pontuação classificatória da anomalia. A incorporação desta distribuição pode acontecer no módulo interno da classificação ou na saída da função de aprendizado. A utilização desta distribuição em módulos internos pode ser encontrada em um estudo do método de aprendizado por reforço baseado na inversa Bayesiana [29], onde a intuição principal é que o comportamento de um agente alimentado com dados sequenciais pode ser aprendido através da sua função de recompensa, de forma que, quando este for alimentado com dados sequenciais anômalos, o resultado desta função poderá ser utilizado para a identificação da anomalia.

Por outro lado, a ideia de impor uma destribuição nos resultados do aprendizado [33] foi motivada, principalmente, pelos comportamentos dos dados vistos em outro estudo, os quais apareciam na forma de Gaussianas [25], levando a aplicação desta antes da codificação da pontuação classificatória da anomalia, viabilizando sua otimização. De modo geral o processo sugere utilizar a proximidade do identificador da anomalia com a média de uma distribuição Gaussiana dos dados proposta através da otimização da rede criada com este fim, utilizando a distância em relação à média em função de um desvio padrão da distribuição anterior como ponto de classificação da anomalia.

As principais vantagens destes métodos são a disponibilização de uma forma de otimização direta da pontuação classificatória da anomalia, diferentemente dos métodos anteriores que tratam de otimizações nos tratamentos dos dados ou dos processos decisórios da classificação, além de facilitar a introdução de outros métodos Bayesianos de aprendizado profundo e viabilizar uma melhor interpretação das identificações realizadas quando se consideram as distribuições selecionadas para aplicação no modelo. A desvantagem, por sua vez, está justamente na definição desta distribuição, visto que é praticamente impossível a identificação de uma que se adeque a todos os dados do mundo real, de forma que a solução acaba sendo atrelada ao evento estudado.

2.1.4.3 Modelos de Máxima Verossimilhança

A abordagem utilizada nestes métodos busca o aprendizado da pontuação classificatória da anomalia pela maximização da verossimilhança. Como instâncias normais e anômalas dos

dados correspondem a padrões frequentes e raros da informação, do ponto de vista probabilístico, presume-se que os eventos normais possuam alta probabilidade, enquanto os anômalos, possuam baixa. Dessa forma, o valor negativo da verossimilhança seria, naturalmente, definido como a pontuação classificatória da anomalia. Em especial, modelos baseados na máxima verossimilhança tem se mostrado efetivos e eficientes com a utilização de ferramentas como a estimativa contrastiva de ruído [13].

A ideia de aprendizado da pontuação classificatória da anomalia através da modelagem da máxima verossimilhança foi introduzida já com a ressalva sobre a sua alta complexidade computacional de tratamento do modelo e o fato de seu foco ser a busca de anomalias em dados categóricos [4]. Motivado por este último ponto, em busca de generalização, o modelo foi adaptado para identificar eventos anômalos utilizando grafos heterogêneos [12], construindo uma estrutura com Perceptrons de múltiplos níveis e um termo de segunda ordem calculado através de um processo com auto-codificadores, atuando simultaneamente, para modelar a interação entre os pares de dados considerados na análise.

A vantagem desta abordagem está na possibilidade da incorporação de novos modelos de interação entre os pares de dados avaliados para adicionar robustez à avaliação do modelo, além de viabilizar a otimização direta da pontuação classificatória da anomalia através da interação entre os dados que queremos avaliar. As desvantagens, por sua vez, ficam por conta da complexidade computacional para a otimização do modelo e de sua dependência da qualidade da combinação dos pares de dados para a geração de valores negativos de verossimilhança a serem avaliados e tratados no treinamento do modelo.

2.1.4.4 Classificação Fim-a-Fim de Classe Única

Esta categoria de métodos busca realizar o treinamento de um classificador capaz de aprender a identificar se um dado é normal ou não de uma forma fim-a-fim, que essencialmente significa que não existem processos intermediários de tratamento dos dados separando os processo de classificação e medição da anomalia. Podemos considerar que métodos semelhantes já foram abordados quando tratamos dos métodos de aprendizado fim-a-fim com pontuação de anomalias (Seção 2.1.4) ou nos métodos utilizando Redes Generativas Adversariais (Seção 2.1.2.2), no entanto, esta abordagem apresenta algumas diferenças chave entre estes pontos considerados nos items acima.

A ideia principal neste método é o aprendizado de um classificador de única classe destinado a identificar as instâncias de dados normais, tal que este irá ser otimizado para classificar esta informação contra dados pseudo-anômalos gerados a partir de Redes Generativas Adversariais, de forma que, nesta abordagem a rede adversarial é usada para otimizar o classificador que está discriminando os dados normais daqueles pseudo-anômalos gerados por ela. Além disso, a pontuação classificatória da anomalia na abordagem da Seção 2.1.2.2 é dada como o resíduo entre o dado real e o dado gerado pela rede, enquanto nesta abordagem o discriminador mencionado acima é aquele utilizado na classificação.

A desvantagem deste método é que não existe uma garantia na captura de anomalias ainda desconhecidas, visto que não se pode afirmar que a rede utilizada para a geração das pseudo-anomalias foi capaz de generalizar de forma satisfatória todo o conjunto de possíveis anomalias. Além disso, ele também pode sofrer com a instabilidade dos modelos baseados em

Redes Generativas Adversariais como mencionado na Seção 2.1.2.2 e ser restrito a cenários semi-supervisionados de detecção de anomalias.

2.2 Identificação de Fraudes

Na seção anterior, diversos estudos tratavam da redução de dimensionalidade ou de tipos específicos de dados, como vídeos e imagens, de forma que podem não ser aplicáveis diretamente no processo de identificação de anomalias em dados financeiros, visto que os cenários que observamos neste caracterizam-se por:

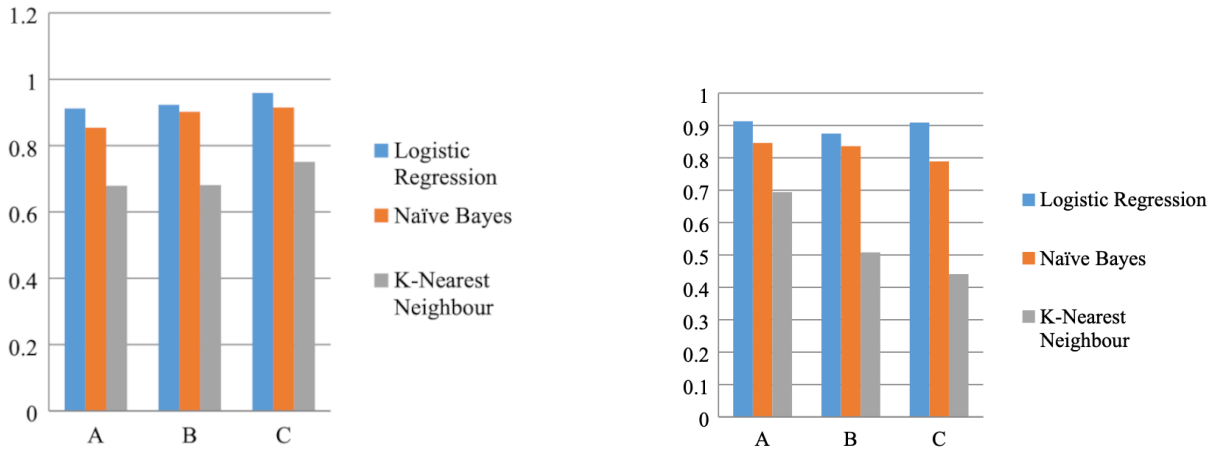
- Dados desbalanceados, com poucas referências de anomalias em comparação ao montante de dados.
- Dados de baixa dimensionalidade visto que na maioria das bases disponíveis não encontramos grandes vetores de características identificando cada transação.
- Dados tabulares, derivados de transações de usuários.

Dessa forma, considerando as bases de dados encontradas na literatura [18, 19, 21, 22], valendo-nos da classificação proposta e guiando-nos pelas abordagems mencionadas nos estudos revistos até o momento que tratavam também de dados tabulares e ofereciam uma boa performance quando analisados cenários de baixa dimensionalidade, buscamos abordagens focadas em identificação de fraudes financeiras.

2.2.1 Identificação de Fraude Utilizando Regressão Logística

Itoo et al. [17] propõem uma comparação entre alguns métodos de aprendizado supervisionado e não supervisionado para a detecção de anomalias em transações de cartão de crédito. Dentre eles, encontra-se a utilização de regressão logística para a caracterização de transações fraudulentas, o que parece uma ideia promissora quando consideramos a premissa de que o comportamento de uma fraude é conhecido e que temos dados suficientes para a extração de características que as identifiquem, sendo uma alternativa para o início da avaliação dos cenários disponíveis inclusive para a análise de impacto do desbalanceamento das bases, algo que pode representar importante complicador aos métodos não supervisionados.

Considerando a base proposta por Itoo et al. [18], os autores realizaram um comparativo da eficácia entre os três algoritmos, demonstrando a melhor eficiência da Regressão Logística, em relação ao KNN e ao Naïve Bayes (vide Figura 2.2).



(a) Acurácia **(b)** F1

Figura 2.2: Resultados apresentados para acurácia e F1 na base proposta por Itoo et al. [18].

2.2.2 Identificação de Fraudes com Rede de Desvios

Pang et al. [33] consideram a identificação fraudes através de um processo fim-a-fim com uma abordagem semi-supervisionada de treinamento de uma Rede de Desvio (*Deviation Network, DevNet*) que utiliza uma distribuição prévia de probabilidades de anomalias para ajustar o aprendizado da pontuação classificatória final da informação submetida à análise.

Esta distribuição considera a probabilidade de um dado conhecido como não-anômalo, de ser uma anomalia e serve como entrada para a função de custo que será utilizada na rede de desvio para a obtenção da pontuação final de classificação do dado sendo avaliado. Dessa forma, os passos realizados no processo são os seguintes:

- Utilização de uma rede de classificação de anomalias para definir um escalar que representa a pontuação classificatória de anomalia para cada dado a ser avaliado. Trata-se aqui de um modelo aprendido e calibrado com a iteração deste processo em um conjunto prévio de treinamento e caracterizado pela função ϕ;
- Para guiar o aprendizado da pontuação classificatória da anomalia, gera-se um segundo valor através de um gerador de números classificatórios de referência, o qual é definido como a média dos elementos $\{r_1, r_2, ..., r_l\}$ criados a partir de um conjunto aleatório l de elementos selecionados de objetos normais e denominado $\mu_{\mathcal{R}}$. Note que este valor em questão pode ser tanto aprendido a partir de uma rede utilizada para a geração dos valores de referência, quanto através de uma distribuição prévia $\mathcal{F}$, método pelo qual o modelo em questão gera as informações desejadas.
- Alimentação da função de perda L com $\phi(x)$, a média $\mu_{\mathcal{R}}$ e seu referido desvio $\sigma_{\mathcal{R}}$ para guiar o aprendizado e otimização da pontuação classificatória final da anomalia.

A Figura 2.3 ilustra uma representação do processo realizado por Pang et al. [33].

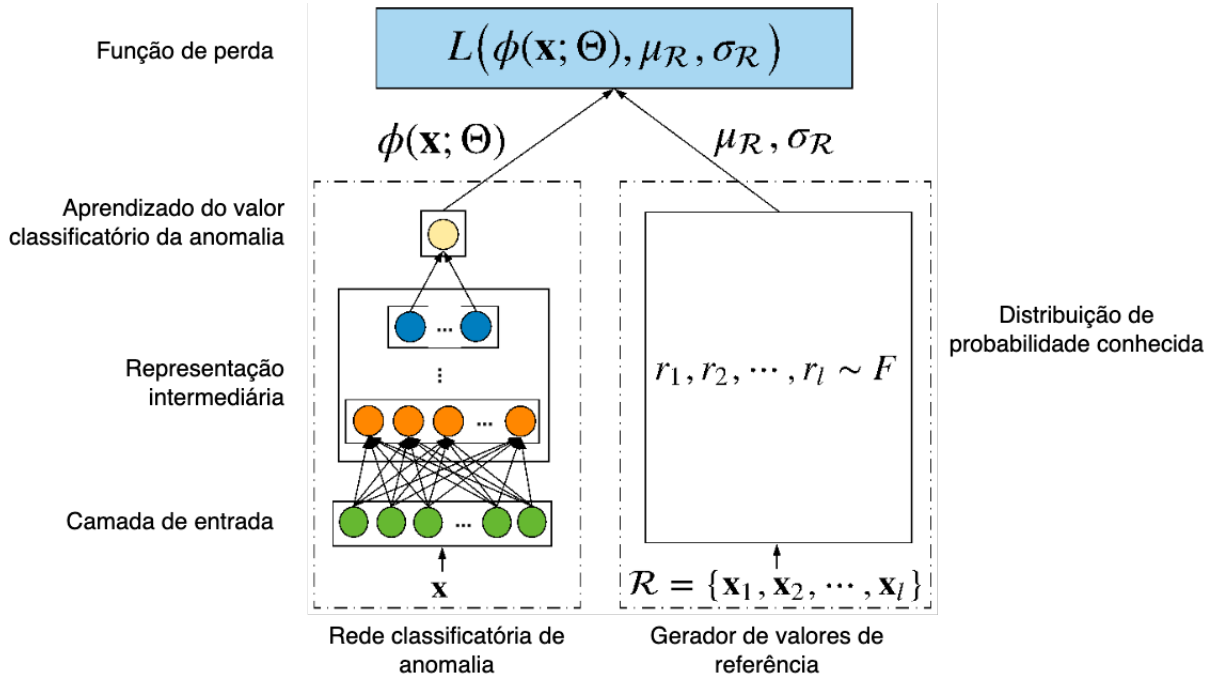


Figura 2.3: Modelo reproduzido a partir do proposto por Pang et al. [33], no qual $\phi(x)$ é um modelo de aprendizado com parâmetro Θ; $\mu_{\mathcal{R}}$ é a média da pontuação classificatória de anomalia de alguns objetos normais, o qual é determinado a partir da distribuição de probabilidades $\mathcal{F}$; $\sigma_{\mathcal{R}}$ é o desvio padrão assossiado à $\mu_{\mathcal{R}}$; a função de custo $\mathcal{L}(\phi(z;\Theta))$ é definida de forma a buscar uma pontuação classificatória da anomalia que possua uma diferença estatisticamente representativa da média $\mu_{\mathcal{R}}$ ao mesmo tempo que seus valores para dados normais aproximem-se de $\mu_{\mathcal{R}}$.

A função de custo é definida através do aprendizado contrastivo, utilizando o desvio entre amostras do conjunto e vetores de características similares, com o objetivo da minimização desta distância [14] de forma que temos, para a rede de desvio:

$$dev(x) = \frac{\phi(x;\Theta) - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}}} \tag{2.12}$$

Substituindo esta na função de custo L:

$$L(\phi(x;\Theta),\mu_{\mathcal{R}},\sigma_{\mathcal{R}}) = (1-y)|dev(x)| + \max(0, a - dev(x))y \tag{2.13}$$

Pelos experimentos reportados pelos autores [33], a abordagem demonstrou um comportamento superior aos comparativos realizados com métodos que utilizam a abordagem que divide em etapas o processo de identificação de uma anomalia, além de demonstrar resultados melhores que os métodos que utilizam apenas o aprendizado não supervisionado, tornando a exploração dessa abordagem, em nossos dados, uma fonte interessante para avaliação.

2.2.3 Identificação de Fraudes com Aumento de Gradiente Extremo

Hajek et al. [15] propõem uma abordagem utilizando *eXtreme Gradient Boosting* (XGBoost), Aumento de Gradiente Extremo, combinada com Subamostragem Aleatória (*Random Under-*

Sampling - RUS), para buscar endereçar elementos do processo de detecção de anomalias, promovendo melhorias na capacidade de tratamento de grandes quantidades de dados em tempo hábil com o XGBoost e no uso destes dados desbalanceados com o RUS.

A abordagem principal utilizada pode ser sumarizada Figura 2.4.

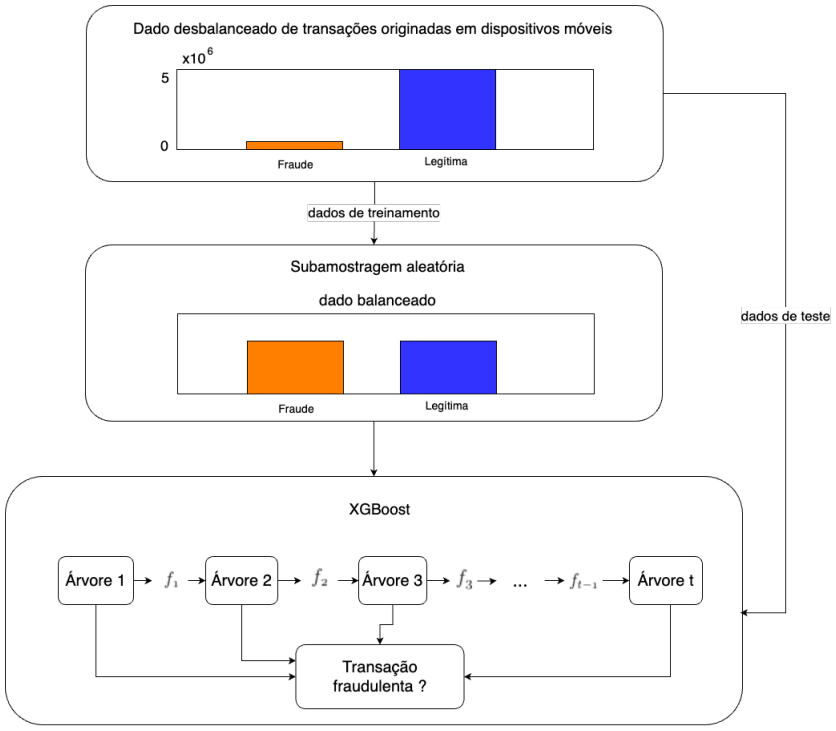


Figura 2.4: Fluxo do Aumento de Gradiente Extremo (XGBoost) com Subamostragem Aleatória (RUS) para a identificação de fraude [15].

A função objetivo do XGBoost a ser otimizada pode ser definida como:

$$obj^{(t)} = \sum_{i=1}^{n}(y_i - (\hat{y}_i^{(t-1)} + f_t(x_i))^2) + \sum_{t=1}^{T}\Omega(f_t) \tag{2.14}$$

Esta abordagem trouxe resultados interessantes demonstrados por Hajek et al. [15], no entanto seu desenvolvimento e validação aconteceram unicamente para a base PaySim [23], de forma que podemos constribuir com a avaliação deste modelo nos conjuntos de dados que estamos considerando em nosso estudo, servindo como base para a busca de otimizações que ajudem na generalização desta abordagem.

2.2.4 Identificação de Fraudes com *Transformers*

Transformer é um modelo de aprendizado profundo construído com base no mecanismo de auto-atenção (*self-attention*) e originalmente foi utilizado no processamento de linguagem natural com excelentes resultados, principalmente pela sua capacidade de tratamento de sequências de dados de forma eficiente e com baixa complexidade computacional [39].

Estendendo o comportamento do *Transformer* padrão, temos a abordagem utilizada na construção do que foi chamado pelos seus criadores de DynGraphTrans [44], uma rede neural que utiliza o mecanismo de auto-atenção para a identificação de anomalias em grafos dinâmicos, os quais mudam ao longo do tempo, como é o caso dos dados financeiros que estamos considerando neste estudo.

A abordagem proposta por Zhang et al. [44] utiliza o mecanismo de auto-atenção para a construção de uma rede neural que busca aprender, a partir de um grafo construído com as transações de um cliente, a representação de cada nó do grafo dinâmico, de forma que a partir desta representação seja possível identificar anomalias naquele nó.

Para isso, a rede utiliza uma função de perda que busca minimizar a distância entre a representação do nó e a representação de seus vizinhos, de forma que, quanto mais distante a representação do nó estiver da representação de seus vizinhos, maior será a pontuação classificatória de anomalia atribuída a ele.

A alteração do modelo tradicional de *Transformers* para a utilização nos grafos de transações propostos por Zhang et al. [44] pode ser vista na Figura 2.5.

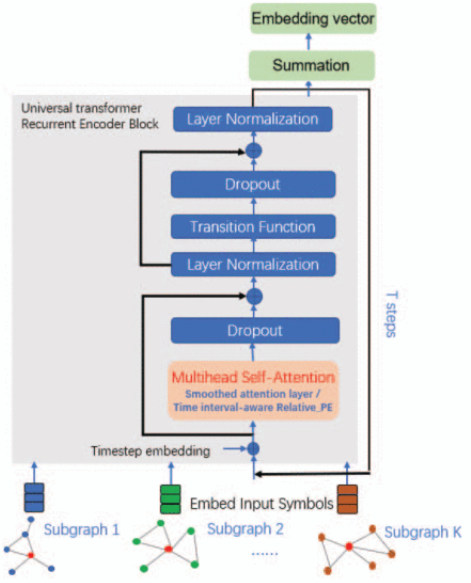


Figura 2.5: Arquitetura do DynGraphTrans [44] com duas modificações nos módulos de auto-atenção em relação à abordagem tradicional de *Transformers*: 1) uma camada extra de atenção suavizada 2) codificação posicional levando em consideração a abordagem de intervalos temporais nos dados.

Um ponto de atenção, neste caso, é a definição de uma janela temporal adequada para a criação dos grafos dinâmicos e o fato que esta abordagem considera a utilização dos dados separados por conta a ser analisada. O que parece interessante quando consideramos que uma anomalia está intimamente ligada às características transacionais de uma conta, mas também um complicador quanto à complexidade computacional de manutenção e cálculo dos modelos para cada conta operacional no sistema.

Esta abordagem trouxe bons resultados, demonstrados por Zhang et al. [44], onde inclusive modelos que faziam uso do Aumento de Gradiente Extremo (XGBoost), como utilizados

por Hajek et al. [15], foram superados, demonstrando ser interessante sua utilização nas bases consideradas em nosso estudo.

Capítulo 3

Metodologia, Material e Métricas

Este capítulo descreve a metodologia, a base de dados, as métricas de avaliação e os recursos computacionais que serão utilizados no desenvolvimento do projeto.

3.1 Metodologia

A metodologia para o desenvolvimento deste projeto consiste no fluxo proposto pela Figura 3.1. As etapas planejadas para resolução do problema serão descritas nas próximas subseções.

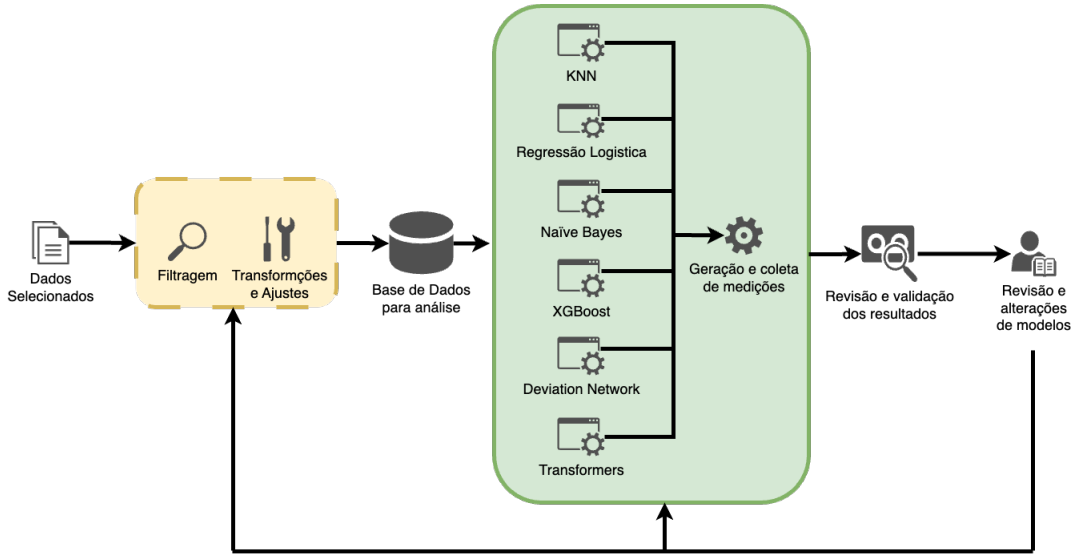


Figura 3.1: Fluxo proposto para resolução do problema.

Inicialmente, serão selecionados os algoritmos de aprendizado de máquina e aprendizado profundo que serão utilizados para a identificação de fraudes. Estes algoritmos serão selecionados com base na revisão bibliográfica realizada na Seção 2.2.

3.1.1 Pré-processamento dos Dados

Após a definição de uma das bases a serem utilizadas, será realizada uma análise exploratória dos dados, com o objetivo de identificar possíveis inconsistências e dados faltantes. A partir desta análise, serão definidas as estratégias de pré-processamento dos dados, que podem incluir a remoção de características, a remoção de registros, a imputação de dados faltantes, a normalização dos dados, a transformação de características categóricas, entre outras.

3.1.2 Processamento dos Dados Selecionados

Após o pré-processamento dos dados, os algoritmos selecionados serão aplicados aos dados, gerando modelos de identificação de fraudes. Estes modelos serão avaliados com base nas métricas definidas na Seção 3.3.

Os dados serão divididos em conjuntos de treinamento, validação e teste. A separação dos dados em conjuntos de treinamento e validação será realizada de forma estratificada, de forma que a proporção de transações fraudulentas e lícitas seja mantida em ambos os conjuntos. O conjunto de validação será utilizado para a seleção dos hiperparâmetros dos algoritmos, de forma a evitar o *overfitting* dos modelos.

A mesma separação de conjuntos será aplicada em todos os modelos, de forma que a viabilizar a comparação entre os resultados encontrados por cada um deles.

3.1.3 Revisão e Validação dos Resultados

Após a geração das métricas propostas na Seção 3.3, será realizada uma revisão dos resultados obtidos, com o objetivo de identificar possíveis inconsistências e erros.

Neste momento, para cada combinação de base de dados e estratégia de pré-processamento, teremos um conjunto de métricas como resultado da aplicação de um algoritmo de identificação de fraudes, de forma que, no cenário onde um destes apresente resultados inconsistentes, a seguinte abordagem será aplicada:

- Caso a maioria dos algoritmos apresente resultados consistentes, os resultados do algoritmo inconsistente serão descartados.
- Caso a maioria dos algoritmos apresente resultados inconsistentes, todas as medições serão descartadas e a estratégia de pré-processamento dos dados será revista.

3.1.4 Revisão das Hipóteses e Alterações de Modelos

Após a revisão de consistência das métricas geradas é possível realizar uma avaliação da eficácia dos modelos selecionados na identificação das fraudes nos diversos conjuntos de dados

utilizados. Com base nessa avaliação, será possível realizar uma revisão das hipóteses levantadas na Seção 1.3 e promover alterações nos modelos selecionados e no pré-processamento dos dados, buscando corroborar as hipóteses consideradas no trabalho.

A partir destas alterações nos modelos e pré-processamento dos dados, será realizada mais uma iteração pelo fluxo completo proposto na Figura 3.1, de forma a validar as alterações e corroborar ou não as hipóteses levantadas.

3.2 Bases de Dados

Para a realização do fluxo proposto na Figura 3.1, será utilizado um conjunto de bases públicas compostas de dados com números variáveis de características, bem como com quantidades diferentes de transações identificadas como fraudulentas, sendo que a semelhança entre todas elas é o desbalanceamento entre a quantidade de objetos marcados como transações legítimas e fraudulentas.

3.2.1 IEEE-CIS Fraud Detection

Base de dados composta por um conjunto de características de transações reais utilizadas em uma competição promovida pela IEEE [22] para o desenvolvimento de novos modelos de identificação de Fraudes.

O conjunto de dados é composto por 590.540 transações, sendo que destas, 20.663 são identificadas como fraudulentas, ou seja, aproximadamente, 3,5% do total. Cada transação é composta por 394 características, sendo que destas, 14 delas são categóricas e 380 numéricas, sendo que uma delas indica o momento da transação e outra, a identificação única desta.

3.2.2 Credit Card Fraud Detection

Este conjunto de dados [18] contém transações feitas por cartões de crédito em setembro de 2013 por titulares de cartões europeus. Este conjunto de dados apresenta transações que ocorreram em dois dias, onde temos 492 fraudes em 284.807 transações. O conjunto de dados é altamente desequilibrado, a classe positiva (fraudes) representa 0,172% de todas as transações.

A base contém apenas valores numéricos que são o resultado de uma transformação de Análise de Componentes Principais (*Principal Component Analysis* - PCA) e devido a restrições de confidencialidade, não são fornecidas as características originais destes dados. As características V1, V2,..., V28 são as componentes principais obtidas com PCA. As únicas características que não foram transformadas são 'Time' e 'Amount'. A característica 'Time' contém os segundos decorridos entre cada transação e a primeira transação no conjunto de dados. A característica 'Amount' é o valor da transação. A característica 'Class' é a variável de resposta e assume o valor 1 em caso de fraude e 0 caso contrário.

3.2.3 Credit Card Transactions Fraud Detection Dataset

Este é um conjunto de dados [20] simulado de transações com cartão de crédito, contendo transações legítimas e fraudulentas no período de 1º de janeiro de 2019 a 31 de dezembro de 2020. Ele abrange cartões de crédito de 1000 clientes realizando transações com um conjunto de 800 comerciantes.

O conjunto de dados é composto por 1.296.675 de transações, sendo que destas, 7.506 são marcadas como fraude. Cada transação é composta por 23 características, das quais 12 são categóricas e 11 numéricas, sendo que destas, 2 são identificadores únicos da transação e do momento em que esta aconteceu.

3.2.4 Synthetic Financial Datasets for Fraud Detection

Este conjunto de dados [23] apresenta transações sintéticas realizadas a partir de dispositivos móveis. Conjuntos de dados financeiros como estes são importantes para muitos pesquisadores, especialmente para aqueles realizando pesquisas no campo da detecção de fraudes. Parte do problema é a natureza intrinsecamente privada das transações financeiras, o que resulta na ausência de conjuntos de dados publicamente disponíveis.

Esta base apresenta um conjunto de dados sintéticos gerados usando um simulador chamado PaySim [27] como uma abordagem para contribuir na disponibilização de bases de teste para o estudo de fraudes no sistema de transações financeiras. O PaySim utiliza dados agregados de um conjunto privado para gerar um conjunto de dados sintético que se assemelha à operação normal de transações e injeta comportamento malicioso para avaliar posteriormente o desempenho de métodos de detecção de fraudes.

O PaySim simula transações realizadas a partir de dispositivos móveis com base em amostras reais extraídas de 30 dias de registros financeiros de um serviço de pagamentos móveis implementado em um país africano, resultando em 6.362.620 transações, sendo que destas, 8.213 são marcadas como fraude. Cada transação é composta por 9 características, das quais 3 são categóricas e 6 numéricas.

Os registros originais foram fornecidos por uma empresa multinacional, que é a provedora do serviço financeiro móvel de forma que maiores detalhes sobre estes dados foram ocultados por questões de confidencialidade.

3.3 Métricas de Avaliação

No processo de identificação de fraudes, o objetivo será sempre maximizar os cenários de acerto, ou seja, aqueles no qual o algoritmo identifica uma transação fraudulenta com sucesso, e minimizar os cenários nos quais uma transação não é identificada como fraudulenta pelo algoritmo, quando de fato é fraudulenta, ou é identificada como tal, mas não é fraudulenta.

Considerando as bases de dados indicadas na Seção 3.2 e os algoritmos selecionados na Seção 2.2, as seguintes métricas serão utilizadas para a avaliação dos resultados obtidos e serão base para a busca de extensões que melhorem a performance desses modelos:

- Taxa de falso positivo (*False Positive Rate* ou FPR), que é a proporção de transações lícitas que foram identificadas como fraudulentas pelo algoritmo.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3.1}$$

onde FP é o número de falsos positivos e TN é o número de verdadeiros negativos.

- Precisão, que representa a proporção entre o número de transações corretamente identificadas como fraudulentas versus o número total de transações que deveriam ser marcadas como tal.

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.2}$$

onde TP é o número de verdadeiros positivos e FP é o número de falsos positivos. Esta métrica é preferível no sistema financeiro [15] em detrimento da taxa de positivos verdadeiros dada a dificuldade na identificação dos falsos negativos quando se trata de fraudes financeiras.

- Revocação (*Recall*), que representa a proporção entre o número de transações corretamente identificadas como fraudulentas versus o número total de transações fraudulentas.

$$\text{Revocação} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

- Área sob a curva ROC (AUC), que é a área sob a curva ROC (*Receiver Operating Characteristic*) e representa a probabilidade de que um classificador atribua uma pontuação mais alta a uma transação fraudulenta do que a uma transação lícita.

$$\text{AUC} = \int_0^1 \text{Revocação}(T) \times \frac{\text{d}}{\text{d}T}\text{FPR}(T)\text{d}T \tag{3.4}$$

onde T é o limiar de decisão do classificador.

- Medida F1, que é a média harmônica entre a precisão e a revocação.

$$\text{F1} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.5}$$

Com estas métricas o objetivo é avaliar a eficácia dos modelos selecionados nas bases definidas, compondo uma medida da capacidade de generalização desses modelos nas diversas bases de dados, abrindo caminho para a criação e avaliação de extensões na busca de melhorias destes modelos para o aumento de suas capacidades de identificação de fraudes.

3.4 Recursos Computacionais

A implementação deste projeto será feita em linguagem de programação Python, devido ao grande número de bibliotecas disponíveis e com boa documentação. O projeto utilizará bibliotecas de aprendizado de máquina, aprendizado profundo, funções científicas e numéricas

e apresentação de gráficos. Algumas bibliotecas que podem ser destacadas são: `NumPy`[1], `scikit-learn`[2], `TensorFlow`[3] e `Matplotlib`[4].

Os experimentos deste projeto serão realizados em três ambientes:

- Notebook MacBook Air com processador Apple M1, 8GB de RAM e macOS Ventura 13.4.
- Notebook Lenovo E470 com processador Intel Core i5 e 16GB de RAM.
- Container CentOS com um processador e 2GB de RAM, com auto-scaling ativado na nuvem do Google.

O objetivo é realizar os experimentos menores e validações de conceito utilizando os notebooks, enquanto os processo de treinamento de mais larga escala serão realizados no Google Cloud, a plataforma de nuvem da Google.

[1] `https://www.numpy.org`
[2] `https://scikit-learn.org`
[3] `https://www.tensorflow.org`
[4] `https://matplotlib.org`

Capítulo 4

Resultados Preliminares

Este capítulo apresenta os resultados dos experimentos preliminares realizados.

4.1 Experimentos Iniciais

O primeiro objetivo foi explorar uma abordagem mais simples em uma das bases de dados selecionadas e avaliar a sua eficácia. Para isso, foi selecionada a base de dados *Credit Card Fraud Detection* [18].

A base de dados foi dividida em 70% para treinamento e 30% para teste. A Tabela 4.1 apresenta os resultados obtidos com uma abordagem simples utilizando KNN como proposto por Itoo et al. [17].

Classe	**Precisão**	**Revocação**	**Medida F1**
Legítima	1.00	1.00	1.00
Fraude	0.94	0.74	0.83

Tabela 4.1: Precisão, Revocação e Medida F1 para a base de dados *Credit Card Fraud Detection* utilizando KNN.

Como esperado, considerando a característica da base, possuindo poucos eventos anômalos, o modelo acertou as transações lícitas e sua performance nesta base foi melhor do que previsto por Itoo et al. [17].

No entanto, quando avaliamos a eficácia da Regressão Logística, temos os resultados apresentados na Tabela 4.2:

Classe	**Precisão**	**Revocação**	**Medida F1**
Legítima	1.00	1.00	1.00
Fraude	0.88	0.63	0.74

Tabela 4.2: Precisão, Revocação e F1 score para a base de dados *Credit Card Fraud Detection* utilizando regressão logística.

A partir dos resultados apresentados, pode-se observar uma eficácia pior do que prevista por Itoo et al. [17], indicando a alta variabilidade nos resultados destes modelos e sua alta dependência da manipulação dos dados utilizado, visto que a base considerada para o experimento foi a mesma utilizada por Itoo et al. [17], com a diferença que no cenário que utilizado aqui, não foram realizadas as mesmas manipulações de dados aplicadas pelos autores.

Estes resultados preliminares de uma análise simples de algoritmos conhecidos em uma base diferente daquela utilizada por Itoo et al. mostram a necessidade da continuidade da busca por um modelo capaz de obter uma eficácia com relação à generalização de seus resultados visto que apenas abrindo mão do pré-processamento realizado por Itoo et al. [17] na base de dados *Credit Card Fraud Detection* [18] já foi possível observar uma variabilidade de resultados que coloca em cheque a avaliação da eficácia dos algoritmos em questão.

Capítulo 5

Plano de Trabalho e Cronograma de Execução

O plano de trabalho é composto pelas seguintes atividades:

1. Estudo e análise das principais técnicas e abordagens disponíveis na literatura.
2. Preparação da base de dados.
3. Construção dos experimentos com as bases públicas identificadas.
4. Realização de testes.
5. Comparação de resultados obtidos entre técnicas e bases de dados.
6. Documentação e publicação dos resultados.
7. Escrita da dissertação.
8. Defesa da dissertação.

O cronograma de execução das atividades propostas, divididas em 4 etapas, em um prazo de 24 meses, é apresentado na Tabela 5.1.

Atividades	**1º ano**						**2º ano**					
	1	**2**	**3**	**4**	**5**	**6**	**1**	**2**	**3**	**4**	**5**	**6**
Etapa 1 - Preparação												
Disciplinas de Pós-Graduação	•	•	•	•	•	•						
Pesquisa bibliográfica	•	•	•	•	•	•	•	•	•			
Preparação da base de dados							•	•	•			
Etapa 2 - Implementações												
Implementação dos algoritmos							•	•	•			
Treinamento e validação dos modelos							•	•	•			
Verificação dos resultados								•	•	•		
Etapa 3 - Testes												
Aplicação dos modelos em múltiplas bases								•	•	•		
Revisão dos resultados								•	•	•		
Comparação com outros trabalhos								•	•	•		
Etapa 4 - Conclusão												
Publicação dos resultados										•	•	
Escrita da dissertação										•	•	•
Defesa da dissertação												•

Tabela 5.1: Cronograma de atividades dividido em bimestres.

Bibliografia

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In *Asian Conference on Computer Vision (ACCV)*, pages 622–637, 2018.

[2] J. Andrews, T. Tanay, E. J. Morton, and L. D. Griffin. Transfer representation-learning for anomaly detection. In *International Conference in Machine Learning (JMLR)*, volume 48, pages 1–5, 2016.

[3] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[4] T. Chen, L. Tang, Y. Sun, Z. Chen, and K. Zhang. Entity Embedding-based Anomaly Detection for Heterogeneous Categorical Events. *arXiv:1608.07502*, pages 1–8, 2016.

[5] B. C. do Brasil. Banco Central do Brasil. `https://www.bcb.gov.br/`. Accessed: 2023-06-04.

[6] B. C. do Brasil. Fintechs. `https://www.bcb.gov.br/estabilidadefinanceira/fintechs`. Accessed: 2023-06-04.

[7] B. C. do Brasil. Manual de Tempos do Pix. `https://www.bcb.gov.br/content/estabilidadefinanceira/pix/Regulamento_Pix/versoes_futuras/IX_ManualdeTemposdoPix-versao5-0.pdf`. Accessed: 2023-06-04.

[8] B. C. do Brasil. O que é Pix? `https://www.bcb.gov.br/estabilidadefinanceira/pix`. Accessed at 2023-06-04.

[9] B. C. do Brasil. TED, DOC e book transfer: entenda como funcionam os tipos de transferências entre contas. `https://www.bcb.gov.br/detalhenoticia/327/noticia`. Accessed: 2023-06-04.

[10] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial Feature Learning. *arXiv:1605.09782*, pages 1–18, 2016.

[11] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134, 2016.

[12] S. Fan, C. Shi, and X. Wang. Abnormal event detection via heterogeneous information network embedding. In *ACM International Conference on Information and Knowledge Management (CIKM)*, page 1483–1486. ACM, 2018.

[13] M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 9, pages 297–304. PMLR, 2010.

[14] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1735–1742, 2006.

[15] P. Hajek, M. Z. Abedin, and U. Sivarajah. Fraud detection in mobile payment systems using an XGBoost-based framework. *Information Systems Frontiers*, 2022.

[16] S. Hawkins, H. He, G. Williams, and R. Baxter. Outlier detection using replicator neural networks. In *International Conference on Data Warehousing and Knowledge Discovery (DaWaK)*, pages 170–180. Springer, 2002.

[17] F. Itoo, Meenakshi, and S. Singh. Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection. *International Journal of Information Technology*, 13:1503–1511, 2021.

[18] Kaggle. Credit Card Fraud Detection. `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud`. Accessed at 2023-07-16.

[19] Kaggle. Credit Card Fraud Detection. `https://www.kaggle.com/datasets/mishra5001/credit-card`. Accessed at 2023-07-16.

[20] Kaggle. Credit Card Transactions Fraud Detection Dataset. `https://www.kaggle.com/datasets/kartik2112/fraud-detection`. Accessed at 2023-07-16.

[21] Kaggle. Fraud Detection - Credit Card. `https://www.kaggle.com/datasets/yashpaloswal/fraud-detection-credit-card`. Accessed at 2023-07-16.

[22] Kaggle. IEEE-CIS Fraud Detection. `https://www.kaggle.com/competitions/ieee-fraud-detection/data`. Accessed at 2023-07-16.

[23] Kaggle. Synthetic Financial Datasets For Fraud Detection. `https://www.kaggle.com/datasets/ealaxi/paysim1`. Accessed at 2023-09-25.

[24] T. Kieu, B. Yang, C. Guo, and C. S. Jensen. Outlier detection for time series with recurrent autoencoder ensembles. In *International Joint Conference of Artificial Intelligence (IJCAI)*, pages 2725–2732, 2019.

[25] H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek. Interpreting and Unifying Outlier Scores. In *SIAM International Conference on Data Mining (SDM)*, pages 13–24, 2011.

[26] B. Liao, J. Zhang, C. Wu, D. McIlwraith, T. Chen, S. Yang, Y. Guo, and F. Wu. Deep sequence learning with auxiliary information for traffic prediction. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 537–546, 2018.

[27] A. Lopez-Rojas, Elmir. Paysim A financial mobile money simulator for fraud detection. In *European Modeling and Simulation Symposium (EMSS)*, pages 249–255, 2016.

[28] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv:1611.02163*, pages 1–18, 2016.

[29] M.-h. Oh and G. Iyengar. Sequential anomaly detection using inverse reinforcement learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1480–1490, 2019.

[30] G. Pang, L. Cao, L. Chen, and H. Liu. Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2041–2050, 2018.

[31] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. Deep learning for anomaly detection: A review. *ACM Compututing Surveys*, 54(2):1–38, 2021.

[32] G. Pang, C. Shen, H. Jin, and A. van den Hengel. Deep weakly-supervised anomaly detection. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1795–1807, 2023.

[33] G. Pang, C. Shen, and A. van den Hengel. Deep anomaly detection with deviation networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 353–362, 2019.

[34] G. Pang, C. Yan, C. Shen, A. v. d. Hengel, and X. Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12173–12182, 2020.

[35] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft. Deep Semi-Supervised Anomaly Detection. *arXiv:1906.02694*, pages 1–23, 2019.

[36] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[37] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27:1–9, 2014.

[38] D. M. Tax and R. P. Duin. Support vector data description. *Machine Learning*, 54:45–66, 2004.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.

[40] H. Wang, G. Pang, C. Shen, and C. Ma. Unsupervised Representation Learning by Predicting Random Distances. *arXiv:1912.12186*, pages 1–18, 2019.

[41] Y. B. Yann LeCun and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[42] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar. Efficient GAN-Based Anomaly Detection. *arXiv:1802.06222*, pages 1–13, 2018.

[43] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar. Adversarially learned anomaly detection. In *IEEE International Conference on Data Mining (ICDM)*, pages 727–736. IEEE, 2018.

[44] S. Zhang, T. Suzumura, and L. Zhang. Dyngraphtrans: Dynamic Graph Embedding via Modified Universal Transformer Networks for Financial Transaction Data. In *IEEE International Conference on Smart Data Services (SMDS)*, pages 184–191, 2021.

[45] C. Zhou and R. C. Paffenroth. Anomaly detection with robust deep autoencoders. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 665–674, 2017.

[46] Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.