UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA DO MARANHÃO
ENGENHARIA DA COMPUTAÇÃO

PATRICK MORAES CUTRIM

Estratégias de Balanceamento de dados em Machine Learning: uma análise interclasse e intraclasse

São Luís - MA
2026

PATRICK MORAES CUTRIM

Estratégias de Balanceamento de dados em Machine Learning: uma análise interclasse e intraclasse

Trabalho de Conclusão do Curso, como requisito obrigatório para a obtenção do título de Bacharel em Engenhraria da Computação pela da Universidade Federal do Maranhão (UFMA), Campus Cidade Universitária Dom Delgado - São Luís, MA.

Orientador: Prof. Dr. Bruno Feres de Souza

São Luís - MA

2026

Cutrim, Patrick Moraes
Estratégias de Balanceamento de dados em Machine Learning: uma análise interclasse e intraclasse / PATRICK MORAES CUTRIM. – São Luís - MA, 2026-

68p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Bruno Feres de Souza
Monografia (Graduação) – , 2026.

1. Aprendizado de Máquina. 2. Desbalanceamento Intraclasse. 3. Desbalanceamento Interclasse. I. Orientador. II. Universidade Federal do Maranhão. III. Centro de Ciências Exatas e Tecnologia. IV. Título

PATRICK MORAES CUTRIM

Estratégias de Balanceamento de dados em Machine Learning: uma análise interclasse e intraclasse

Trabalho de Conclusão do Curso, como requisito obrigatório para a obtenção do título de Bacharel em Engenhraria da Computação pela da Universidade Federal do Maranhão (UFMA), Campus Cidade Universitária Dom Delgado - São Luís, MA.

Trabalho aprovado em São Luís - MA, 20 de julho de 2026

Prof. Dr. Bruno Feres de Souza
Orientador

Prof. Dr. Paulo Rogerio de Almeida Ribeiro
Examinador

Prof. Dr. Pedro Baptista Fernandes
Examinador

São Luís - MA
2026

Dedico este trabalho aos meus pais, às minhas avós e, em memória, ao meu avô, que sempre foram minha inspiração e força ao longo dessa caminhada..

Agradecimentos

Agradeço, primeiramente, a Deus, por ter sido meu amparo e minha força ao longo de toda esta caminhada, iluminando meus passos nos momentos de dúvida e me sustentando diante de cada desafio. À minha família, meu porto seguro, dedico a mais profunda gratidão. Aos meus pais, pelo amor incondicional, pela educação que me deram e por jamais medirem esforços para que eu chegasse até aqui. Às minhas avós e, em memória, ao meu avô, agradeço por serem inspiração e exemplo de vida. Sem o apoio, o incentivo e a compreensão de vocês, este trabalho não teria sido possível. Ao meu orientador, Prof. Dr. Bruno Feres de Souza, expresso minha sincera gratidão pela dedicação, paciência e generosidade em compartilhar seus conhecimentos. As orientações precisas e o acompanhamento atento foram fundamentais para o desenvolvimento e a conclusão desta pesquisa. Aos meus amigos, agradeço pela companhia, pelo apoio nos momentos difíceis e por tornarem essa jornada mais leve. As conversas, o incentivo e a parceria ao longo desses anos deixaram marcas que levarei para toda a vida. A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho, o meu muito obrigado.

"Uma mente que se abre a uma nova ideia
jamais voltará ao seu tamanho original."

Albert Einstein

Resumo

O desbalanceamento de dados constitui um desafio fundamental no campo do aprendizado de máquina, afetando diretamente a confiabilidade de modelos preditivos em domínios sensíveis como o setor financeiro e a saúde. Embora o desequilíbrio interclasse tenha sido o foco histórico das pesquisas, o fenômeno do desbalanceamento intraclasse — caracterizado pela distribuição desigual de padrões e níveis de dificuldade no interior de uma mesma categoria — apresenta-se como um obstáculo crítico à capacidade de generalização dos algoritmos. Este trabalho explora estratégias de mitigação para ambos os níveis de desbalanceamento, conduzindo uma análise experimental em dois domínios de alta criticidade: a detecção de fraudes em transações de cartão de crédito e a classificação de tumores cerebrais em imagens de ressonância magnética. A metodologia adota uma abordagem bimodal: para os dados tabulares, empregam-se técnicas de reamostragem híbrida e adaptativa, como SMOTE-ENN e ADASYN, combinadas a modelos de *Gradient Boosting* de alto desempenho (XGBoost, LightGBM e CatBoost); para as imagens médicas, recorre-se a arquiteturas com mecanismos de atenção, incluindo a EfficientNet-B0 com CBAM e o Swin Transformer. A eficácia das abordagens é auditada por meio de um protocolo de validação cruzada estratificada, fundamentado em métricas adequadas a distribuições assimétricas, como F1-Score, AUC-ROC e G-Mean. Os resultados demonstram que a incorporação da diversidade interna das classes ao processo de treinamento resulta em modelos mais robustos: no domínio financeiro, os algoritmos de *Gradient Boosting* alcançaram desempenho equivalente, com G-Mean de 0,9245 para o CatBoost, enquanto no diagnóstico oncológico o Swin Transformer atingiu G-Mean de 0,9773 e acurácia de 97,74%. Conclui-se que o tratamento explícito do desbalanceamento intraclasse é determinante para a identificação confiável de subgrupos minoritários, reforçando a robustez e a equidade preditiva dos sistemas. **Palavras-chave**:

Aprendizado de Máquina. Desbalanceamento Intraclasse. Desbalanceamento Interclasse.

Abstract

Data imbalance constitutes a fundamental challenge in machine learning, directly affecting the reliability of predictive models in sensitive domains such as the financial sector and healthcare. Although inter-class imbalance has historically been the focus of research, the phenomenon of intra-class imbalance — characterized by the uneven distribution of patterns and difficulty levels within the same category — poses a critical obstacle to the generalization capacity of algorithms. This work explores mitigation strategies for both levels of imbalance, conducting an experimental analysis in two high-criticality domains: fraud detection in credit card transactions and brain tumor classification in magnetic resonance images. The methodology adopts a bimodal approach: for tabular data, hybrid and adaptive resampling techniques such as SMOTE-ENN and ADASYN are employed, combined with high-performance *Gradient Boosting* models (XGBoost, LightGBM, and CatBoost); for medical images, attention-based architectures are used, including EfficientNet-B0 with CBAM and the Swin Transformer. The effectiveness of the approaches is audited through a stratified cross-validation protocol, grounded in performance metrics suitable for asymmetric distributions, such as F1-Score, AUC-ROC, and G-Mean. The results demonstrate that incorporating the internal diversity of classes into the training process yields more robust models: in the financial domain, the *Gradient Boosting* algorithms achieved equivalent performance, with a G-Mean of 0.9245 for CatBoost, while in the oncological diagnosis task the Swin Transformer reached a G-Mean of 0.9773 and an accuracy of 97.74%. We conclude that explicitly addressing intra-class imbalance is decisive for the reliable identification of minority subgroups, reinforcing the robustness and predictive fairness of the systems.

Keywords: Machine Learning. Intra-class Imbalance. Inter-class Imbalance.

Lista de ilustrações

Figura 1 – Exemplo de desequilíbrio intraclasse. 14
Figura 2 – Exemplo de desequilíbrio interclasse. 15
Figura 3 – Exemplo de árvore de decisão para detecção de fraudes. 20
Figura 4 – Exemplo de arquitetura de rede neural convolucional simples. 21
Figura 5 – Arquitetura do EfficientNet-b0 com integração do módulo CBAM. . . . 22
Figura 6 – Arquitetura hierárquica do Swin Transformer com janelas deslizantes. . 22
Figura 7 – Amostras representativas do *Brain Tumor MRI Dataset*, exibindo as quatro categorias em diferentes planos de aquisição. 37
Figura 8 – Distribuição original dos dados (Desbalanceamento Extremo). 49
Figura 9 – Matriz de Correlação dos Atributos Transformados (PCA). 50
Figura 10 – Matriz de Confusão: Modelo CatBoost + SMOTE-ENN. 52
Figura 11 – Curva ROC Multiclasse: Desempenho do Swin Transformer. 54
Figura 12 – Matriz de Confusão: Classificação Multiclasse de Tumores Cerebrais. . 56

Lista de tabelas

Tabela 1 – Matriz de Confusão: representação dos erros e acertos de um modelo de classificação. 25
Tabela 2 – Cenários indicados para uso das métricas em dados desbalanceados . . 27
Tabela 3 – Desempenho dos Algoritmos de Boosting com SMOTE-ENN (Média $k = 5$) 52
Tabela 4 – Comparativo de desempenho das arquiteturas no conjunto de teste — Base MRI (média macro). 54
Tabela 5 – Relatório de classificação multiclasse — Swin Transformer. 55
Tabela 6 – Impacto das técnicas de reamostragem no espaço latente — Base MRI. 57
Tabela 7 – Comparação de G-Mean entre estratégias de balanceamento (Média $k = 5$) 59

Lista de abreviaturas e siglas

ADASYN	*Adaptive Synthetic Sampling*
ANN	Redes Neurais Artificiais (*Artificial Neural Networks*)
AUC	Área Sob a Curva (*Area Under the Curve*)
CBAM	*Convolutional Block Attention Module*
CNN	Redes Neurais Convolucionais (*Convolutional Neural Networks*)
EFB	*Exclusive Feature Bundling*
ENN	*Edited Nearest Neighbors*
G-Mean	Média Geométrica (*Geometric Mean*)
GOSS	*Gradient-based One-Side Sampling*
HEM	*Hard Example Mining*
IA	Inteligência Artificial
LSTM	*Long Short-Term Memory*
MAE	Erro Médio Absoluto (*Mean Absolute Error*)
ML	Aprendizado de Máquina (*Machine Learning*)
MRI	Ressonância Magnética (*Magnetic Resonance Imaging*)
MSE	Erro Quadrático Médio (*Mean Squared Error*)
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
RNA	Redes Neurais Artificiais
ROC	Característica de Operação do Receptor (*Receiver Operating Characteristic*)
RMSE	Raiz do Erro Quadrático Médio (*Root Mean Square Error*)
SMOTE	*Synthetic Minority Over-sampling Technique*
UFMA	Universidade Federal do Maranhão

Sumário

1 INTRODUÇÃO 14
1.1 Justificativa 16
1.2 Objetivos 16
1.2.1 Objetivo Geral 16
1.2.2 Objetivos Específicos 16
1.3 Estrutura do Trabalho 17

2 FUNDAMENTAÇÃO TEÓRICA 18
2.1 Conceitos de Aprendizado de Máquina 18
2.1.1 Tipos de Aprendizado de Máquina 18
2.1.2 Algoritmos de Classificação 19
2.1.2.1 Árvore de Decisão 19
2.1.2.2 Redes Neurais Convolucionais (CNN) 20
2.1.2.3 EfficientNet com CBAM 21
2.1.2.4 Modelos Baseados em Gradient Boosting 23
2.2 Métricas de Avaliação para Dados Desbalanceados 24
2.2.1 Matriz de Confusão 25
2.2.2 Precisão, Recall, F1-Score, Média Geométrica e AUC-ROC 25

3 TÉCNICAS DE DESBALANCEAMENTO DE DADOS 28
3.1 Linha de base sem tratamento 29
3.2 Reamostragem aleatória 29
3.2.1 Undersampling aleatório 29
3.2.2 Oversampling aleatório 30
3.3 SMOTE-ENN 30
3.4 ADASYN 31
3.5 Hard Example Mining (HEM) 31

4 METODOLOGIA 34
4.1 Caracterização e Seleção dos Corpora de Dados 34
4.1.1 Dataset de Fraude em Cartão de Crédito (ULB) 35
4.1.2 Dataset de Tumores Cerebrais (MRI) 35
4.2 Tratamento do Desbalanceamento: Estratégias de Reamostragem Híbrida 37
4.2.1 ADASYN (Adaptive Synthetic Sampling) 38
4.3 Arquiteturas de Modelagem Preditiva e Paradigmas de Aprendizado 40

4.3.1 Modelos de Gradient Boosting para Dados Tabulares . . . 40
4.3.2 Mecanismos de Atenção para Extração de Características Discriminativas . 41
4.3.2.1 CBAM: Convolutional Block Attention Module . . . 41
4.3.2.1.1 1. Channel Attention Module (CAM) . . . 41
4.3.2.1.2 2. Spatial Attention Module (SAM) . . . 42
4.3.2.2 Swin Transformer: Auto-Atenção Hierárquica com Janelas Deslocadas . . . 42
4.3.2.2.1 Auto-Atenção em Janelas Locais . . . 42
4.3.2.2.2 Mecanismo de Janelas Deslocadas (Shifted Windows) . . . 43
4.3.2.2.3 Aplicação ao Diagnóstico de Tumores . . . 43
4.4 Protocolo de Validação e Métricas de Performance . . . 44

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS . . . 48
5.0.1 A Falácia da Acurácia Nominal . . . 48
5.1 Resultados no Domínio Financeiro: Detecção de Fraudes . . . 49
5.1.1 Diagnóstico e Impacto da Reamostragem Híbrida (SMOTE-ENN) . . . 50
5.1.2 Avaliação Comparativa de Desempenho: O Domínio do CatBoost . . . 51
5.2 Resultados no Domínio da Saúde: Classificação de Tumores Cerebrais 53
5.2.1 Desempenho das Arquiteturas e Impacto da Atenção . . . 53
5.2.2 Análise Quantitativa: Relatório de Classificação . . . 54
5.2.3 Análise Qualitativa: Diagnóstico via Matriz de Confusão . . . 55
5.2.4 Impacto da Reamostragem no Espaço Latente . . . 57
5.3 Discussão Transversal: Convergência Metodológica e o Desafio Intraclasse . . . 58
5.4 Resultados das Estratégias de Desbalanceamento . . . 59
5.5 Comparação de Desempenho entre as Técnicas . . . 59
5.6 Análise das Métricas e Impacto do Desbalanceamento Intraclasse . 60
5.7 Limitações e Possíveis Fontes de Erro . . . 60

6 CONCLUSÃO . . . 62
6.1 Considerações Finais . . . 62
6.2 Contribuições do Trabalho . . . 63
6.3 Sugestões para Trabalhos Futuros . . . 64

REFERÊNCIAS . . . 66

1 Introdução

A maioria dos algoritmos de aprendizado de máquina assume distribuição uniforme entre as classes. Na prática, essa suposição raramente se confirma, pois em aplicações reais observa-se frequentemente o *desbalanceamento de dados*, que ocorre quando a proporção de amostras é significativamente desigual entre classes ou subgrupos, prejudicando a capacidade preditiva do modelo.

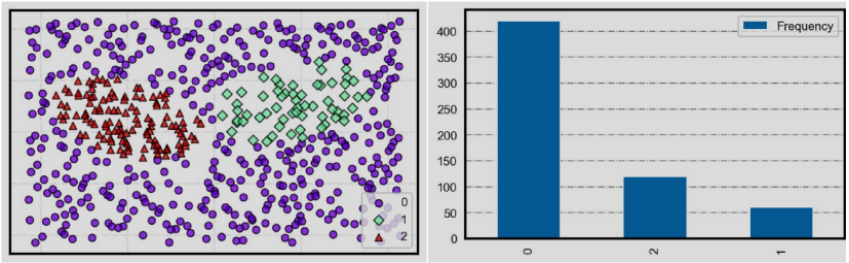
O desbalanceamento afeta domínios sensíveis: em diagnósticos médicos, subtipos raros sub-representados aumentam falsos negativos; em segurança cibernética, ataques com assinaturas pouco frequentes apresentam detecção inferior a 50%; no setor financeiro, fraudes representam menos de 0,2% do total de transações, levando algoritmos a negligenciarem detecção de fraudes.

A literatura concentrou-se, por décadas, no *desequilíbrio interclasse* — disparidade no número de exemplos entre classes distintas. Desenvolveram-se métodos de pré-processamento (sobreamostragem e subamostragem), abordagens baseadas em custo e métodos híbridos. Entretanto, identificou-se uma lacuna relevante: a maioria dos estudos ignora o *desequilíbrio intraclasse*, que ocorre quando há variação significativa na distribuição de exemplos dentro de uma mesma classe.

O *desequilíbrio intraclasse* manifesta-se quando determinados subgrupos caracterizados por padrões específicos encontram-se sub-representados, mesmo pertencendo à mesma categoria. Causas incluem: sobreposição entre fronteiras de decisão, presença de ruído ou existência de subconceitos raros. Nessas condições, modelos tendem a priorizar padrões dominantes, negligenciando instâncias mais difíceis de classificar, comprometendo diagnósticos de subtipos específicos de doenças, reconhecimento facial de minorias étnicas e análise de padrões anômalos.

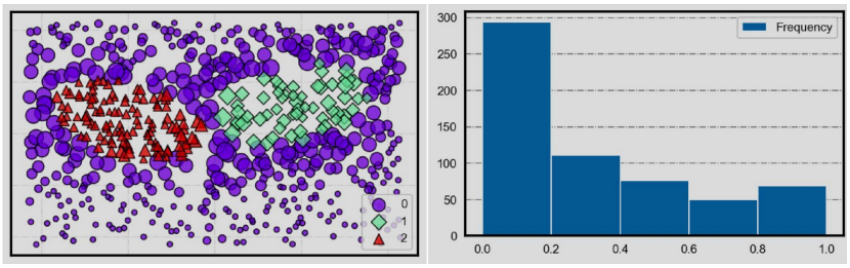
A Figura 1 ilustra um exemplo de distribuição intraclasse desbalanceada, enquanto a Figura 2 apresenta, comparativamente, a distribuição interclasse.

Figura 1 – Exemplo de desequilíbrio intraclasse.



Fonte: Liu et al. (2021).

Figura 2 – Exemplo de desequilíbrio interclasse.



Fonte: Liu et al. (2021).

A Figura 1 ilustra o fenômeno de desequilíbrio intraclasse através da distribuição desigual de exemplos dentro de uma mesma classe. O painel à esquerda mostra a classe majoritária (pontos roxos) dividida em dois subgrupos distintos: um com alta densidade (região central) e outro com baixa densidade (região periférica). O gráfico de barras evidencia que o subgrupo denso representa aproximadamente 70% dos exemplos intraclasse, enquanto o subgrupo esparso representa apenas 30%. A classe minoritária (pontos verdes) complementa a visualização, demonstrando como a complexidade da classificação aumenta com variabilidade significativa dentro de uma única classe.

A Figura 2 apresenta o desequilíbrio interclasse, onde uma classe dominante (pontos roxos/azuis) representa aproximadamente 80% dos dados, enquanto demais classes constituem apenas 20%. Este cenário é típico de detecção de fraudes, onde transações legítimas vastamente superam as fraudulentas. Ambas as figuras ilustram como os dois níveis de desbalanceamento ocorrem simultaneamente em conjuntos reais, exigindo abordagens diferenciadas.

Apesar dos avanços em técnicas para tratamento de dados desbalanceados, poucas abordagens tratam explicitamente o desequilíbrio intraclasse. Métodos desenvolvidos para desbalanceamento interclasse não apresentam bom desempenho em contextos intraclasse, pois não consideram as particularidades de cada subgrupo. Além disso, grande parte das pesquisas foca exclusivamente em redes neurais profundas, sem explorar suficientemente aplicabilidade em outros algoritmos como árvores de decisão ou modelos probabilísticos.

Diante desse cenário, propõe-se investigar e implementar estratégias para lidar com o desequilíbrio intraclasse mantendo foco no desequilíbrio interclasse. Busca-se responder às seguintes questões:

1. De que maneira as diferenças internas a uma mesma classe influenciam o desempenho de modelos de aprendizado de máquina?

2. Quais fatores intrínsecos aos dados contribuem para o surgimento do desequilíbrio intraclasse?

3. É possível desenvolver estratégias que tratem simultaneamente ambos os desequilíbrios, preservando generalização e sensibilidade em subgrupos minoritários?

1.1 Justificativa

A investigação do desequilíbrio intraclasse justifica-se pela necessidade de aprimorar robustez e capacidade de generalização de modelos preditivos, especialmente em domínios nos quais variabilidade interna de uma classe representa fator determinante para precisão. Embora desbalanceamento interclasse tenha recebido ampla atenção, variações internas de uma mesma classe permanecem subexploradas.

Métodos tradicionais como *oversampling* e *undersampling* mostram-se eficazes no balanceamento entre classes, mas não corrigem assimetria entre subgrupos internos. Por exemplo, ao lidar com dados médicos, aumento sintético de casos pode favorecer apenas subtipos frequentes, deixando raros ainda sub-representados. Da mesma forma, redução de amostras pode acarretar perda de informações críticas.

Em aplicações críticas — diagnóstico assistido, segurança pública e detecção de fraudes — a incapacidade de representar adequadamente padrões minoritários pode levar a falhas graves: diagnósticos incorretos, não detecção de ameaças e perda de oportunidades em prevenção. Além disso, desequilíbrio intraclasse pode intensificar viés algorítmico, pois subgrupos menos representados recebem previsões menos precisas, contribuindo para perpetuação de desigualdades em sistemas automatizados.

Dessa forma, a presente pesquisa investiga sistemática e criticamente os impactos e soluções para desequilíbrio intraclasse, contribuindo para avanços em tratamento de dados desbalanceados e oferecendo alternativas metodológicas que melhorem desempenho, ampliem aplicabilidade em contextos reais e fortaleçam confiabilidade.

1.2 Objetivos

1.2.1 Objetivo Geral

Investigar estratégias para lidar com o desequilíbrio intraclasse em aprendizado de máquina, visando melhorar performance de modelos preditivos em cenários de alta complexidade.

1.2.2 Objetivos Específicos

Para alcançar o objetivo geral, este estudo propõe-se a:

- Revisar literatura existente sobre técnicas para tratamento de desbalanceamento, com foco em métodos para desequilíbrio intraclasse;
- Analisar impacto do desequilíbrio intraclasse no desempenho de modelos de aprendizado de máquina;
- Implementar e testar diferentes abordagens para tratá-lo, incluindo técnicas de reamostragem e ajustes no treinamento;
- Avaliar e comparar resultados usando métricas adequadas para cenários desbalanceados;
- Identificar possíveis melhorias e direções futuras para pesquisas na área.

1.3 Estrutura do Trabalho

Este trabalho estrutura-se em seis capítulos: O Capítulo 1 introduz a problemática do desequilíbrio intraclasse, objetivos e justificativa. O Capítulo 2 detalha fundamentação teórica em Aprendizado de Máquina e métricas de performance. O Capítulo 3 explora técnicas de reamostragem e tratamento de desbalanceamento. A metodologia experimental, corpora de dados e infraestrutura encontram-se no Capítulo 4. O Capítulo 5 apresenta análise crítica dos resultados. Por fim, o Capítulo 6 sintetiza conclusões e propõe trabalhos futuros.

2 Fundamentação Teórica

2.1 Conceitos de Aprendizado de Máquina

O aprendizado de máquina (*machine learning* - ML) é uma subárea da inteligência artificial (IA) dedicada ao desenvolvimento de modelos computacionais capazes de identificar padrões e realizar inferências a partir de dados, sem depender de instruções explícitas programadas passo a passo (MITCHELL, 1997; GOODFELLOW; BENGIO; COURVILLE, 2016). Trata-se de um paradigma que evoluiu significativamente desde meados do século XX, impulsionado tanto pelo avanço dos recursos computacionais quanto pela disponibilidade crescente de grandes volumes de dados (*big data*) (DOMINGOS, 2012). A aplicação do aprendizado de máquina abrange uma ampla variedade de domínios, incluindo, mas não se limitando a, visão computacional, bioinformática, análise financeira, segurança cibernética, diagnóstico médico e processamento de linguagem natural (MURPHY, 2012). Seu uso tem se mostrado particularmente eficaz em cenários nos quais padrões complexos e de alta dimensionalidade precisam ser identificados, superando, muitas vezes, a capacidade analítica humana.

No contexto do presente trabalho, compreender os fundamentos do aprendizado de máquina é essencial para a análise das estratégias voltadas ao tratamento de desequilíbrios de dados, sejam eles interclasse ou intraclasse.

2.1.1 Tipos de Aprendizado de Máquina

De acordo com a natureza dos dados disponíveis e a forma como o modelo é treinado, o aprendizado de máquina pode ser categorizado, de forma geral, em três abordagens principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado semi-supervisionado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; BISHOP, 2006).

- **Aprendizado Supervisionado**: Consiste no treinamento de um modelo a partir de um conjunto de dados rotulado, no qual cada instância de entrada está associada a uma saída conhecida (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). O objetivo é que o modelo aprenda a relação entre entradas e saídas, possibilitando a previsão de rótulos para novos dados. Exemplos comuns incluem regressão linear, regressão logística, máquinas de vetores de suporte (SVM), árvores de decisão e redes neurais profundas (GÉRON, 2019). Essa abordagem é amplamente utilizada em tarefas como classificação de imagens, previsão de séries temporais e diagnósticos médicos assistidos.

- **Aprendizado Não Supervisionado**: Nesta abordagem, o modelo recebe apenas dados de entrada, sem rótulos ou saídas conhecidas. O objetivo é identificar estruturas ou padrões ocultos nos dados, como agrupamentos (*clustering*) ou representações compactas (*dimensionality reduction*) (BISHOP, 2006). Técnicas comuns incluem o algoritmo k-means, DBSCAN e análise de componentes principais (PCA). É frequentemente utilizado em segmentação de clientes, detecção de anomalias e compressão de dados.

- **Aprendizado Semi-Supervisionado**: Combina elementos das duas abordagens anteriores, sendo especialmente útil quando se dispõe de uma pequena quantidade de dados rotulados e um grande volume de dados não rotulados (ZHU; GOLDBERG, 2009). O conhecimento extraído do conjunto rotulado orienta o aprendizado sobre o conjunto não rotulado, reduzindo custos e tempo de rotulagem. Essa técnica é aplicada em áreas como reconhecimento de fala, análise de texto e sistemas de recomendação.

O entendimento dessas categorias é fundamental para a escolha das estratégias adequadas no tratamento de dados desbalanceados, visto que cada abordagem apresenta desafios e vantagens específicos quando exposta a distribuições de dados desiguais.

2.1.2 Algoritmos de Classificação

Os algoritmos de classificação desempenham papel central no aprendizado de máquina, sendo amplamente utilizados para a resolução de problemas supervisionados, inclusive em cenários com dados desbalanceados (FERNáNDEZ et al., 2018a). A escolha do algoritmo mais adequado depende não apenas da natureza dos dados, mas também do tipo de tarefa, dos recursos computacionais disponíveis e da complexidade dos padrões que se deseja aprender.

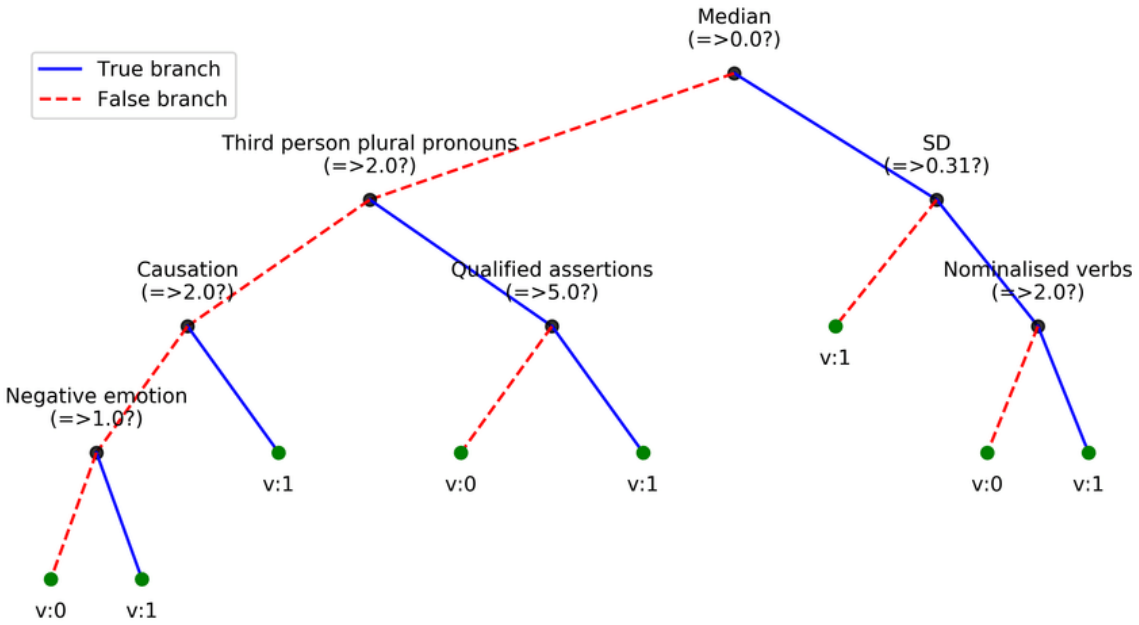
Neste trabalho, optou-se por adotar diferentes abordagens de classificação para dois domínios distintos: detecção de fraudes em transações financeiras e classificação de tumores cerebrais a partir de imagens de ressonância magnética (MRI).

2.1.2.1 Árvore de Decisão

A árvore de decisão é um modelo baseado em uma estrutura hierárquica que divide o espaço de atributos em regiões progressivamente menores, tomando decisões com base em regras derivadas dos dados (QUINLAN, 1986). Cada nó interno representa um teste sobre um atributo, cada ramo corresponde a um resultado possível desse teste e cada nó folha define uma classe final.

Para o problema de detecção de fraudes, a árvore de decisão foi escolhida por sua interpretabilidade, baixo custo computacional e capacidade de lidar com atributos categóricos e numéricos. Além disso, apresenta bom desempenho em bases tabulares e permite a análise clara das regras de decisão, aspecto fundamental em aplicações financeiras onde a transparência do modelo é requisito crítico. A figura abaixo representa uma arvore de decisão para detecção de fraudes.

Figura 3 – Exemplo de árvore de decisão para detecção de fraudes.



Fonte: ResearchGate (2019).

2.1.2.2 Redes Neurais Convolucionais (CNN)

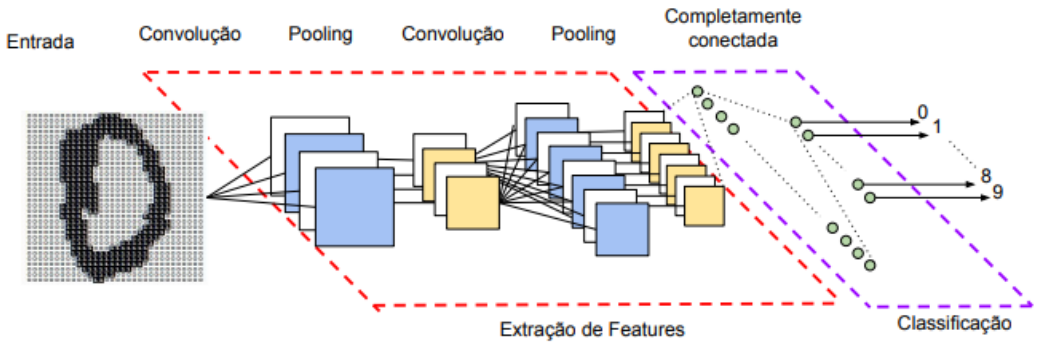
As redes neurais convolucionais (*Convolutional Neural Networks* — CNN) são arquiteturas especializadas no processamento de dados com estrutura de grade, como imagens, utilizando camadas convolucionais para extrair automaticamente características relevantes (LECUN; BENGIO; HINTON, 2015). Neste estudo, para o problema de classificação de tumores cerebrais, empregaram-se três variações principais:

- **CNN Simples**: A rede neural convolucional simples (CNN Simples) é uma arquitetura de profundidade reduzida, composta por um número limitado de camadas convolucionais e de pooling. Essa configuração foi utilizada como linha de base para a avaliação de desempenho, servindo como ponto de comparação para arquiteturas mais complexas, como EfficientNet com CBAM e Swin Transformer.

  A CNN Simples é capaz de extrair características relevantes das imagens por meio das camadas convolucionais, que aplicam filtros sobre a entrada, e das camadas de pooling, que reduzem a dimensionalidade mantendo as informações mais relevantes.

Por fim, as camadas totalmente conectadas realizam a classificação final com base nas características extraídas.

Figura 4 – Exemplo de arquitetura de rede neural convolucional simples.

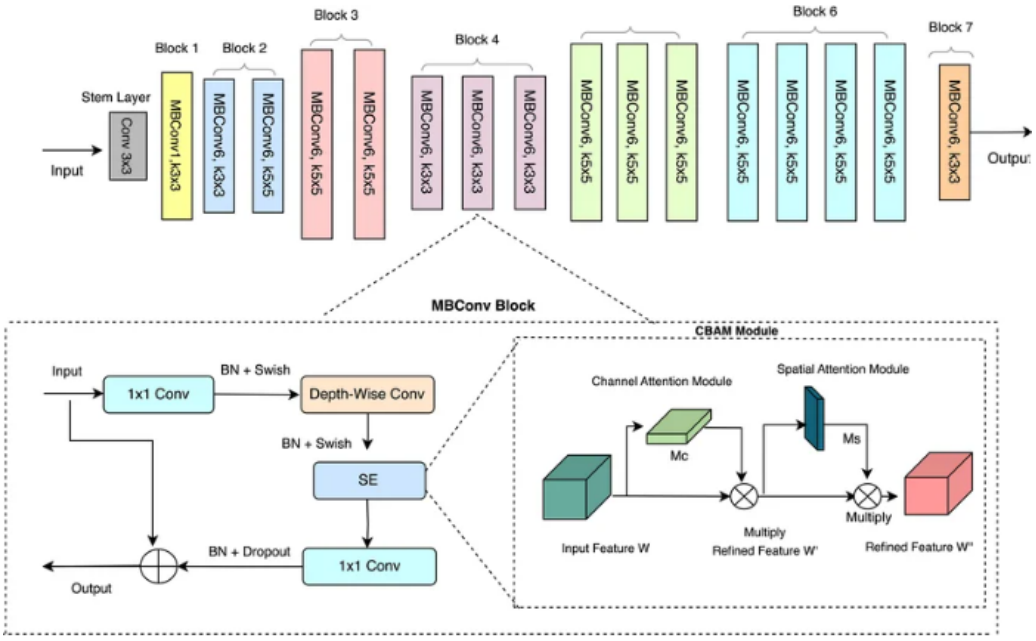


Fonte: Vargas et al. (2016).

2.1.2.3 EfficientNet com CBAM

O EfficientNet é uma família de modelos convolucionais que otimiza simultaneamente a largura, profundidade e resolução da rede de forma equilibrada, alcançando alto desempenho com menor custo computacional (TAN; LE, 2019). Neste trabalho, integrou-se o módulo CBAM (*Convolutional Block Attention Module*) (WOO et al., 2018), que adiciona mecanismos de atenção espacial e de canal à arquitetura. O CBAM permite que a rede concentre-se em regiões e características mais relevantes das imagens, aprimorando a capacidade de extração de padrões discriminativos, o que é particularmente útil em tarefas de classificação de imagens médicas, como a detecção de câncer em imagens de ressonância magnética (MRI).

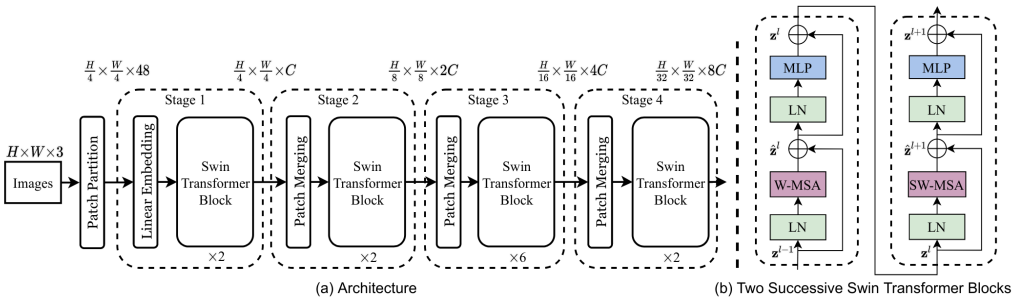
Figura 5 – Arquitetura do EfficientNet-b0 com integração do módulo CBAM.



Fonte: Ding et al. (2022).

- **Swin Transformer**: Arquitetura derivada do conceito de *Vision Transformers* (ViTs), projetada para capturar dependências de longo alcance em imagens por meio de mecanismos de autoatenção. O diferencial do Swin Transformer está na utilização de uma abordagem hierárquica com janelas deslizantes (*shifted windows*), que possibilita a modelagem eficiente de relações locais e globais, reduzindo o custo computacional em comparação aos ViTs convencionais. Essa característica o torna adequado para aplicações que exigem análise multiescala, como o processamento de imagens de ressonância magnética (MRI), onde padrões sutis podem ocorrer em diferentes resoluções. A Figura **??** apresenta a arquitetura do Swin Transformer, destacando suas quatro etapas principais (estágios), cada uma composta por blocos sucessivos de atenção com mesclagem progressiva de *patches*, permitindo a extração de características em diferentes níveis de abstração (LIU et al., 2021).

Figura 6 – Arquitetura hierárquica do Swin Transformer com janelas deslizantes.



Fonte: Liu et al. (2021).

A escolha dessas arquiteturas para o problema de câncer visa comparar o desempenho entre um modelo de referência mais simples e arquiteturas de ponta, avaliando o impacto de diferentes níveis de complexidade e mecanismos de atenção na detecção de padrões sutis presentes nas imagens.

2.1.2.4 Modelos Baseados em Gradient Boosting

Os modelos fundamentados em *Gradient Boosting* representam o estado da arte para a modelagem de dados estruturados, operando sob o paradigma de aprendizado por conjunto (*ensemble learning*) de natureza aditiva (CHEN; GUESTRIN, 2016a). Diferente de métodos como o *Random Forest*, o *Boosting* é um processo sequencial desenhado para a redução sistemática do viés (*bias*) através da combinação de classificadores fracos (*weak learners*), tipicamente árvores de decisão de profundidade limitada (CHEN; GUESTRIN, 2016a; KE et al., 2017a).

Matematicamente, o algoritmo procura aproximar uma função $f(x)$ que minimize uma função de perda global $L(y, f(x))$. A cada iteração m, um novo estimador $h_m(x)$ é incorporado para corrigir os resíduos (erros) da etapa anterior, conforme a equação:

$$f_m(x) = f_{m-1}(x) + \eta h_m(x) \tag{2.1}$$

Onde η (*learning rate*) controla a taxa de convergência e auxilia na prevenção do *overfitting* (CHEN; GUESTRIN, 2016a).

Em cenários de desbalanceamento intraclasse, esta técnica demonstra robustez superior pois instâncias de difícil classificação geram resíduos elevados, forçando as árvores subsequentes a especializarem-se nessas regiões de baixa densidade (HE; GARCIA, 2009a). A Figura **??** ilustra uma dessas unidades básicas de decisão aplicada ao contexto deste trabalho.

Neste estudo, implementam-se as três variantes mais eficientes da literatura (CHEN; GUESTRIN, 2016a; KE et al., 2017a; DOROGUSH; ERSHOV; GULIN, 2018a):

- **XGBoost (*eXtreme Gradient Boosting*):** Implementa um *framework* escalável de *tree boosting* que utiliza a expansão de Taylor de segunda ordem na função objetivo para uma aproximação refinada do gradiente, garantindo convergência superior em relação a métodos de primeira ordem (CHEN; GUESTRIN, 2016a). Além de incorporar regularizações $L1$ (*Lasso*) e $L2$ (*Ridge*) para controlar a complexidade estrutural das árvores e prevenir o *overfitting*, o algoritmo introduz um sistema de *sparsity-aware* que permite o tratamento nativo e eficiente de dados esparsos ou com valores ausentes (CHEN; GUESTRIN, 2016a). Sua arquitetura paralela e distribuída o torna ideal para processar os volumosos registros de transações financeiras analisados neste trabalho (CHEN; GUESTRIN, 2016a).

– **LightGBM (*Light Gradient Boosting Machine*):** Diferencia-se pelo emprego de uma estratégia de crescimento de árvore do tipo *Leaf-wise* (por folha) com limitação de profundidade, em contraste com a abordagem *Level-wise* convencional, o que permite uma redução mais acelerada da função de perda (KE et al., 2017a). O modelo utiliza técnicas inovadoras como o *Gradient-based One-Side Sampling* (GOSS) para filtrar instâncias com gradientes pequenos e o *Exclusive Feature Bundling* (EFB) para agrupar atributos esparsos e mutuamente exclusivos (KE et al., 2017a). Tais mecanismos reduzem drasticamente o custo computacional e o consumo de memória, facilitando a análise de padrões complexos em grandes corpora de dados (KE et al., 2017a).

– **CatBoost:** Destaca-se pelo tratamento nativo e sofisticado de atributos categóricos, utilizando esquemas de permutação aleatória para transformar categorias em valores numéricos sem a necessidade de codificações manuais prévias (*one-hot encoding*) (DOROGUSH; ERSHOV; GULIN, 2018a). O algoritmo emprega o conceito de *Ordered Boosting* para combater o viés de predição e o vazamento de informações do alvo (*target leakage*) comuns em métodos de *boosting* tradicionais (DOROGUSH; ERSHOV; GULIN, 2018a). Além disso, a utilização de árvores simétricas (*Symmetric Trees*) confere ao modelo uma robustez superior e acelera significativamente o tempo de inferência, característica crítica para sistemas de detecção em tempo real (DOROGUSH; ERSHOV; GULIN, 2018a).

2.2 Métricas de Avaliação para Dados Desbalanceados

A avaliação de modelos de aprendizado de máquina em cenários com dados desbalanceados exige métricas que representem adequadamente o desempenho preditivo, indo além da simples acurácia (BRANCO; TORGO; RIBEIRO, 2016; FERNáNDEZ et al., 2018a). Em contextos onde há grande disparidade na distribuição das classes, a acurácia tende a superestimar o desempenho, pois um classificador que sempre prediz a classe majoritária pode obter valores elevados dessa métrica, mas falhar completamente em identificar a classe minoritária, que muitas vezes representa o interesse principal (POWERS, 2011).

Para superar essa limitação, utilizam-se métricas baseadas na *matriz de confusão*, capazes de capturar tanto a capacidade do modelo de detectar instâncias positivas quanto a proporção de predições corretas entre as classificadas como positivas. Neste trabalho, são empregadas as seguintes métricas: Precisão (*Precision*), Revocação (*Recall*), F1-Score e Área sob a Curva ROC (AUC-ROC). A escolha dessas métricas justifica-se pela necessidade de avaliar não apenas o acerto geral, mas também a

qualidade da detecção da classe minoritária em cenários críticos, como detecção de fraudes e diagnóstico médico.

2.2.1 Matriz de Confusão

A matriz de confusão organiza as predições do modelo em quatro categorias, permitindo análise detalhada dos acertos e erros de classificação (PROVOST; FAWCETT; KOHAVI, 1998):

	Classe Predita Positiva	**Classe Predita Negativa**
Classe Real Positiva	TP (Verdadeiro Positivo)	FN (Falso Negativo)
Classe Real Negativa	FP (Falso Positivo)	TN (Verdadeiro Negativo)

Tabela 1 – Matriz de Confusão: representação dos erros e acertos de um modelo de classificação.

Onde:

- **Verdadeiro Positivo (TP)**: Instâncias da classe positiva corretamente identificadas.
- **Falso Positivo (FP)**: Instâncias da classe negativa incorretamente classificadas como positivas.
- **Verdadeiro Negativo (TN)**: Instâncias da classe negativa corretamente identificadas.
- **Falso Negativo (FN)**: Instâncias da classe positiva incorretamente classificadas como negativas.

2.2.2 Precisão, Recall, F1-Score, Média Geométrica e AUC-ROC

t As principais métricas adotadas neste trabalho, além da acurácia, são:

- **Precisão (Precision)** Mede a proporção de predições positivas corretas, refletindo a confiabilidade do modelo quando este prediz a classe positiva:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.2}$$

  Alta precisão indica baixo número de falsos positivos, fator crucial em cenários como detecção de fraudes, onde classificações incorretas podem gerar custos significativos.

- **Recall (Sensibilidade ou Revocação)** Avalia a capacidade do modelo de identificar corretamente todos os exemplos positivos:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.3}$$

Um alto recall significa que poucos casos positivos são perdidos, algo essencial em diagnósticos médicos, nos quais falsos negativos podem ter consequências graves.

- **F1-Score** Combina precisão e recall em uma média harmônica, equilibrando as duas métricas:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

O F1-Score é especialmente útil quando há necessidade de balancear a importância entre evitar falsos positivos e falsos negativos.

- **Área sob a Curva ROC (AUC-ROC)** Mede a capacidade geral do modelo de distinguir entre classes positivas e negativas ao longo de diferentes limiares de decisão (FAWCETT, 2006):

$$\text{AUC-ROC} = \int_0^1 TPR\, d(FPR) \tag{2.5}$$

Onde:

  * $TPR = \frac{TP}{TP+FN}$ (Taxa de Verdadeiros Positivos)
  * $FPR = \frac{FP}{FP+TN}$ (Taxa de Falsos Positivos)

Valores de AUC próximos de 1 indicam alto poder discriminativo, enquanto valores próximos de 0,5 indicam desempenho similar a classificações aleatórias.

- $TPR = \frac{TP}{TP+FN}$ (Taxa de Verdadeiros Positivos)
- $FPR = \frac{FP}{FP+TN}$ (Taxa de Falsos Positivos) Valores de AUC próximos de 1 indicam alto poder discriminativo, enquanto valores próximos de 0,5 indicam desempenho similar a classificações aleatórias.
- **G-Mean (Média Geométrica)** Métrica que quantifica o equilíbrio entre a performance nas classes majoritárias e minoritárias simultaneamente, penalizando modelos que negligenciam a classe com menor representatividade:

$$G\text{-}Mean = \sqrt{\text{Recall} \times \text{Specificity}} \tag{2.6}$$

O G-Mean é fundamental para avaliar o equilíbrio entre a sensibilidade nas classes majoritárias e minoritárias de forma concomitante, sendo essencial para garantir que o modelo não seja enviesado pela classe dominante.

Além das definições apresentadas, é importante compreender em quais cenários cada métrica é mais adequada, especialmente quando se lida com dados desbalanceados, onde a acurácia isolada pode ser enganosa. Por exemplo, um modelo que classifica todos os casos como pertencentes à classe majoritária pode atingir uma acurácia elevada, mas falhar completamente na detecção de instâncias da classe minoritária,

tornando-se ineficaz para o problema real (BRANCO; TORGO; RIBEIRO, 2016; FERNáNDEZ et al., 2018a).

A Tabela 2 apresenta um resumo das métricas abordadas e indica em quais contextos cada uma é mais apropriada, bem como exemplos de aplicações típicas.

Tabela 2 – Cenários indicados para uso das métricas em dados desbalanceados

Métrica	**Quando Utilizar**	**Aplicações Típicas**
Precisão (Precision)	Quando é mais importante minimizar falsos positivos, garantindo que predições positivas sejam confiáveis.	Detecção de fraudes financeiras, classificação de e-mails como spam.
Recall (Sensibilidade)	Quando é crítico identificar todos os casos positivos, mesmo que ocorram falsos positivos.	Diagnóstico médico, detecção de doenças raras.
F1-Score	Quando é necessário equilibrar precisão e recall, especialmente em contextos onde há impacto tanto de falsos positivos quanto de falsos negativos.	Processamento de linguagem natural, sistemas de recomendação.
AUC-ROC	Quando se deseja medir a capacidade geral de separação entre classes ao longo de diferentes limiares de decisão.	Avaliação global de classificadores binários, problemas de classificação em aprendizado supervisionado.

Justificativa da escolha

Embora a acurácia seja intuitiva e amplamente utilizada, ela não é confiável em bases de dados desbalanceadas, pois não penaliza adequadamente a má performance na classe minoritária. Neste trabalho, a utilização de Precisão, Recall, F1-Score e AUC-ROC visa fornecer uma visão mais abrangente do desempenho, garantindo que tanto a capacidade de detectar positivos quanto a de evitar erros críticos sejam consideradas de forma equilibrada.

3
Técnicas de Desbalanceamento de Dados

O desbalanceamento intraclasse caracteriza-se pela distribuição desigual de exemplos fáceis e difíceis dentro de uma mesma classe, fenômeno que pode ocorrer mesmo quando o número total de instâncias entre classes está equilibrado (BUDA; MAKI; MAZUROWSKI, 2018; LIU; WEI; TIAN, 2022). Essa assimetria interna tende a surgir em contextos onde há sobreposição de fronteiras de decisão, presença de ruídos nos dados, existência de subgrupos minoritários dentro da classe ou heterogeneidade de padrões e variabilidade intra-classe (MOUSAVI; GHASEMIAN; SHAMSOLMOALI, 2020).

Na prática, isso significa que alguns subconjuntos da classe podem ser mais representativos e fáceis de classificar, enquanto outros, mais raros ou complexos, acabam sendo sub-representados durante o processo de aprendizado. Modelos de aprendizado de máquina, especialmente aqueles que não incorporam mecanismos explícitos de balanceamento ou ponderação, tendem a priorizar padrões mais frequentes e de fácil ajuste, negligenciando regiões menos densas do espaço de atributos. Esse viés de aprendizado resulta em redução da capacidade de generalização, principalmente quando o objetivo é identificar instâncias pertencentes a tais subgrupos minoritários, impactando negativamente o desempenho em tarefas que demandam alta sensibilidade para padrões raros — como detecção de anomalias, diagnósticos médicos e identificação de fraudes. Do ponto de vista geométrico, o desbalanceamento intraclasse afeta a topologia do espaço de decisão: regiões mais densas se expandem e dominam a fronteira, enquanto áreas esparsas sofrem compressão ou são incorretamente englobadas pela classe oposta. Esse efeito é agravado em cenários de alta dimensionalidade, onde a distância entre exemplos se torna menos discriminativa e os subgrupos minoritários passam a estar cercados por instâncias de outras classes, dificultando a separação.

Considerando a relevância desse problema e seu impacto direto na capacidade discriminativa dos modelos, este trabalho adota um conjunto de técnicas para mitigá-lo, abrangendo desde abordagens simples até métodos adaptativos mais robustos. As técnicas selecionadas incluem:

1. **Baseline sem tratamento** — utilizada como linha de referência para medir o impacto das demais técnicas e avaliar o desempenho do modelo sob a distribuição original dos dados.

2. **Reamostragem aleatória** — englobando *undersampling* (redução de exemplos da classe majoritária) e *oversampling* (replicação de exemplos da classe minoritária), como formas diretas de modificar a proporção entre subgrupos.
3. **SMOTE-ENN** — técnica híbrida que combina a geração sintética de amostras minoritárias via SMOTE com a limpeza de ruído próximo à fronteira de decisão por meio do Edited Nearest Neighbors.
4. **ADASYN** — método adaptativo que concentra a geração de exemplos sintéticos em regiões minoritárias de maior dificuldade, ampliando o poder discriminativo em áreas críticas.
5. **Hard Example Mining (HEM)** — estratégia que prioriza exemplos de maior dificuldade durante o treinamento, fortalecendo a capacidade do modelo em aprender padrões complexos ou pouco frequentes.

A adoção dessas técnicas tem como objetivo não apenas melhorar métricas globais de desempenho, mas também **aumentar a cobertura e a sensibilidade para subgrupos específicos**, reduzindo vieses internos e ampliando a robustez do modelo. Nos experimentos conduzidos, essas abordagens foram aplicadas de forma controlada e comparativa, de modo a mensurar não apenas ganhos absolutos, mas também potenciais *trade-offs*, como aumento de falsos positivos ou sobreajuste em regiões de baixa densidade.

3.1 Linha de base sem tratamento

Adotou-se uma linha de base sem qualquer forma de balanceamento (None) para mensurar o desempenho de referência do pipeline de modelagem. Esta condição estabelece um patamar mínimo para comparação e permite avaliar o ganho incremental introduzido por cada técnica de desbalanceamento. No caso de fraudes (dados tabulares), tal baseline permite verificar o quanto a árvore de decisão responde à distribuição original. Para imagens de ressonância magnética (MRI), a linha de base com redes neurais convolucionais evidencia o efeito do desequilíbrio sobre a sensibilidade a subtipos e padrões sutis.

3.2 Reamostragem aleatória

3.2.1 Undersampling aleatório

No undersampling aleatório, é reduzido o número de instâncias da classe majoritária por seleção aleatória, buscando uma razão de equilíbrio definida (*sampling_strategy*) (FERNáNDEZ et al., 2018a). Tal procedimento simplifica a fronteira de decisão

e diminui o custo computacional, porém pode descartar exemplos informativos e empobrecer a diversidade da classe majoritária. Em dados de fraude é recomendado aplicá-lo de forma conservadora (por exemplo, reduzindo gradualmente a razão majoritária) e sempre no conjunto de treino dentro de cada dobra de validação, para evitar vazamento de informação.

3.2.2 Oversampling aleatório

O oversampling aleatório consiste em aumentar artificialmente a representatividade da classe minoritária por meio da replicação direta de instâncias já existentes, selecionadas de forma aleatória, até que seja atingida a proporção desejada entre as classes (FERNáNDEZ et al., 2018a). Essa abordagem é considerada uma das formas mais simples de balanceamento, pois não requer cálculos complexos ou geração de novos exemplos sintéticos. Sua principal vantagem está na implementação rápida e de baixo custo computacional, sendo aplicável em praticamente qualquer tipo de dado. No entanto, a técnica apresenta limitações relevantes. A replicação exata de amostras tende a aumentar o risco de sobreajuste (*overfitting*), já que o modelo pode memorizar padrões específicos dessas instâncias replicadas, em vez de aprender características mais gerais da classe minoritária. Esse efeito é particularmente crítico em conjuntos de dados pequenos, nos quais a duplicação de poucos exemplos pode gerar redundância excessiva no espaço de atributos.

Em cenários com dados tabulares de fraude, o oversampling aleatório foi empregado como uma etapa preliminar, preparando o conjunto para abordagens mais sofisticadas, como SMOTE ou ADASYN, que geram exemplos sintéticos e reduzem a redundância. No caso de imagens médicas de ressonância magnética (MRI), a replicação direta no espaço de pixels não foi utilizada, pois pode introduzir padrões artificiais e não acrescentar variabilidade relevante para o aprendizado. Quando necessário, a operação foi realizada no espaço latente — após a extração de *embeddings* por redes neurais pré-treinadas — ou substituída por técnicas específicas de aumento de dados (*data augmentation*) adequadas ao domínio, como rotações leves, variação de contraste e adição de ruído fisiologicamente plausível.

3.3 SMOTE-ENN

O SMOTE-ENN combina a geração sintética de exemplos minoritários, realizada pelo SMOTE, com a remoção posterior de instâncias ruidosas ou conflitantes por meio do Edited Nearest Neighbors (ENN) (CHAWLA et al., 2002a; BATISTA; PRATI; MONARD, 2004; WILSON, 1972). No SMOTE, para cada instância minoritária x_i, são selecionados k vizinhos minoritários mais próximos e gerados pontos sintéticos

por interpolação linear no segmento entre x_i e um vizinho $x_{i,nn}$:

$$\tilde{x} = x_i + \lambda \cdot (x_{i,nn} - x_i), \quad \lambda \sim \mathcal{U}(0,1).$$

Após essa etapa, o ENN atua como filtro de ruído, removendo instâncias — geralmente da classe majoritária, mas também sintéticas — cujos rótulos divergem da maioria de seus vizinhos, ajustando assim a fronteira de decisão (WILSON, 1972).

Esse encadeamento oferece duas vantagens principais: ampliação controlada da região minoritária e depuração da fronteira de decisão. Em bases tabulares de fraude, é importante ajustar o valor de k (por exemplo, $k \in \{3, 5, 7\}$) e o parâmetro *sampling_strategy* para evitar a geração excessiva de instâncias em regiões muito esparsas. No caso de imagens, não é recomendada a aplicação direta do SMOTE no espaço de pixels, pois isso pode distorcer características relevantes. A alternativa mais adequada é realizar o processo no espaço latente, utilizando *embeddings* extraídos por uma CNN pré-treinada, preservando melhor a semântica de alto nível.

3.4 ADASYN

O ADASYN (Adaptive Synthetic Sampling) concentra a geração de exemplos sintéticos em regiões minoritárias de maior dificuldade, determinadas pela densidade relativa de vizinhos majoritários (HE et al., 2008). Para cada instância minoritária x_i, calcula-se a razão local de dificuldade $r_i = \frac{\Delta_i}{k}$, em que Δ_i é o número de vizinhos majoritários entre os k vizinhos mais próximos. Em seguida, define-se a fração normalizada $\bar{r}_i = \frac{r_i}{\sum_j r_j}$ e a quantidade de amostras a gerar para x_i, dada por $g_i = \bar{r}_i \cdot G$, onde G representa o total desejado de instâncias sintéticas. Os novos exemplos são produzidos de forma semelhante ao SMOTE, mas com intensidade guiada por g_i. Essa estratégia prioriza a expansão de regiões mais complexas, o que tende a melhorar o *recall* da classe minoritária. Entretanto, pode aumentar o risco de amplificar ruído e causar sobreposição entre classes em áreas particularmente confusas.

3.5 Hard Example Mining (HEM)

O Hard Example Mining (HEM) consiste em priorizar, durante o treinamento, instâncias que o modelo encontra maior dificuldade para classificar, identificadas por perdas (*loss*) mais elevadas ou erros recorrentes (SHRIVASTAVA; GUPTA; GIRSHICK, 2016). Existem variações offline, que realizam seleção periódica a partir de um conjunto pré-existente, e variações online, como o *Online Hard Example Mining* (OHEM), que fazem a seleção de forma dinâmica dentro de cada minibatch.

Em redes neurais convolucionais aplicadas a exames de ressonância magnética, o HEM pode ser implementado por:

- Aumentar a probabilidade de amostragem de exemplos com maiores perdas em cada época, mantendo também uma proporção mínima de exemplos fáceis para garantir estabilidade;
- Ajustar limiares de confiança por classe, reforçando a exposição a padrões raros intraclasse;
- Monitorar a distribuição de perdas por subgrupos, verificando se a priorização não introduz viés indesejado.

Para os dados tabulares, o *Hard Example Mining* foi implementado por meio da reponderação iterativa das instâncias, conforme o Código 3.1. A cada rodada, os exemplos com maior erro de predição (os 20% mais difíceis) recebem peso ampliado no re-treinamento, forçando o modelo a especializar-se nas regiões de fronteira.

```
def treinar_com_hem(modelo_fn, X_tr, y_tr, rounds=3, boost=2.0):
    w = np.ones(len(y_tr))
    for _ in range(rounds):
        modelo = modelo_fn()
        modelo.fit(X_tr, y_tr, sample_weight=w)
        proba = modelo.predict_proba(X_tr)[:, 1]
        # Dificuldade = erro absoluto entre probabilidade e rÓtulo
        dificuldade = np.abs(proba - y_tr.values)
        limiar = np.quantile(dificuldade, 0.80)   # top 20% mais
            dificeis
        w = np.where(dificuldade > limiar, w * boost, w)
        w = w / w.mean()                           # normaliza os
            pesos
    return modelo
```

Listing 3.1 – Hard Example Mining via reponderação iterativa dos exemplos difíceis.

Em dados tabulares, como no caso de detecção de fraudes com modelos baseados em árvores de decisão, a ideia é implementada por meio da reponderação iterativa de instâncias com maior erro em validações internas (usando, por exemplo, o parâmetro *sample_weight* em ciclos de reentreinamento) ou com *bootstraps* que superamostram os casos mais problemáticos. Em ambos os cenários, é recomendável alternar fases de priorização com fases de treinamento neutro, evitando que a fronteira de decisão se ajuste apenas a regiões excessivamente difíceis.

A escolha das técnicas de balanceamento foi considerada pelo tipo de dado que foi trabalhado, a distribuição das classes e o impacto potencial de cada abordagem sobre métricas críticas como *recall*, precisão e F1-score. Essa avaliação orienta a aplicação de métodos de forma estratégica, preservando a integridade das amostras

e maximizando o desempenho preditivo. A próxima seção apresenta a metodologia experimental utilizada, descrevendo a preparação dos dados, a configuração dos modelos e os procedimentos de avaliação.

4
Metodologia

4.1 Caracterização e Seleção dos Corpora de Dados

Os códigos-fonte completos dos experimentos, incluindo o pré-processamento, o treinamento dos modelos e a geração das métricas, encontram-se disponíveis publicamente[1].

A eficácia de modelos de Aprendizado de Máquina, especialmente em cenários de alta criticidade, é intrinsecamente dependente da qualidade, volume e representatividade estatística dos dados utilizados no treinamento. Em cenários de **desbalanceamento intraclasse**, o desafio transcende a simples disparidade numérica entre categorias (desbalanceamento interclasse); ele reside na presença de subgrupos heterogêneos dentro de uma mesma classe, onde padrões minoritários podem ser erroneamente assimilados como ruído ou outliers pelo classificador (BRANCO; TORGO; RIBEIRO, 2016; HE; GARCIA, 2009b). Para garantir que as estratégias de mitigação propostas nesta pesquisa possuam validade científica e capacidade de generalização, adotou-se uma **abordagem bimodal**. Esta escolha metodológica visa testar os algoritmos em dois extremos do processamento de dados na Engenharia:

1. **Dados Tabulares de Alta Dimensionalidade**: Representados pelo domínio financeiro, onde as variáveis são atributos estruturados, muitas vezes transformados matematicamente (como via PCA), exigindo que o modelo aprenda correlações lineares e não-lineares em espaços latentes (ULB, 2016).
2. **Dados Não Estruturados (Imagens)**: Representados pelo domínio da saúde, onde o modelo deve extrair características morfológicas e espaciais complexas através de tensores, lidando com a variabilidade de textura e intensidade de pixel inerente a exames de imagem médica (BRAIN. . . , 2025).

A seleção desses corpora não foi apenas técnica, mas estratégica. Ao validar as mesmas métricas de performance (como o G-Mean) em domínios tão distintos quanto a detecção de fraudes e o diagnóstico oncológico, o trabalho demonstra que o tratamento do desbalanceamento intraclasse é uma necessidade transversal na Engenharia da Computação. Além disso, a utilização de datasets públicos

[1] Detecção de fraudes: <https://colab.research.google.com/drive/11meRirM-kdchnPAd7lcS0e7I0RrYG_RY?usp=sharing>. Classificação de tumores (MRI): <https://colab.research.google.com/drive/1TLwI0zz9JeNYDgd8ptQge20z27BRdIEg?usp=sharing>.

de referência assegura a **reprodutibilidade experimental**, permitindo que o protocolo de validação cruzada estratificada ($k = 5$) seja auditado e comparado com o estado da arte da literatura acadêmica (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017; FERNáNDEZ et al., 2018b).

4.1.1 Dataset de Fraude em Cartão de Crédito (ULB)

O primeiro corpus selecionado para a experimentação em dados tabulares é o *Credit Card Fraud Detection Dataset*, que reúne registros de transações financeiras efetuadas por portadores de cartões europeus durante o período de setembro de 2013 (ULB, 2016). Composto por uma volumetria massiva de 284.807 instâncias, a base apresenta um cenário de desbalanceamento interclasse extremo, no qual apenas 492 registros (equivalentes a 0,172% do total) são categorizados como fraudulentos, impondo um desafio rigoroso à convergência de algoritmos convencionais que tendem ao viés em favor da classe majoritária (ULB, 2016). No que tange à arquitetura dos atributos, a base é constituída por variáveis preditivas em que a maioria (denominadas de V1 a V28) foi submetida a uma transformação linear via Análise de Componentes Principais (PCA) para assegurar a confidencialidade e o sigilo de dados sensíveis, restando apenas os atributos 'Time' e 'Amount' em seus formatos brutos originais (JOLLIFFE, 2002; ULB, 2016). A complexidade inerente a este *dataset*, contudo, reside fundamentalmente no desbalanceamento intraclasse presente na categoria majoritária; a classe de transações legítimas não constitui um conjunto homogêneo, mas sim uma sobreposição de diversos padrões comportamentais — variando de microtransações cotidianas a vultosas compras sazonais — gerando subgrupos com densidades estatísticas e distribuições de probabilidade distintas que podem obscurecer as assinaturas anômalas das fraudes, exigindo estratégias de modelagem capazes de capturar tais nuances conceituais sem comprometer a especificidade do modelo (BUREZ; POEL, 2009; ULB, 2016).

4.1.2 Dataset de Tumores Cerebrais (MRI)

Para a validação das estratégias no domínio da visão computacional, utilizou-se o *Brain Tumor MRI Dataset*, um corpus robusto constituído por 7.023 imagens de ressonância magnética (MRI) axial, coronais e sagitais, estruturado para o diagnóstico de quatro categorias distintas: glioma, meningioma, tumor de pituitária e a ausência de patologias (no tumor) (BRAIN. . . , 2025). A relevância deste conjunto de dados para o estudo do desbalanceamento intraclasse reside na heterogeneidade morfológica das lesões e, primordialmente, na dispersão estatística da classe de controle; a categoria sem tumor apresenta uma alta variância interna decorrente de diferenças anatômicas interpacientes, variações na angulação da captura e a presença

de artefatos de imagem que podem mimetizar padrões patológicos sutis, desafiando a capacidade de generalização do modelo (BRAIN..., 2025; LIU et al., 2021). Tecnicamente, o processamento destas instâncias exigiu a implementação de um *pipeline* de normalização rigoroso, no qual as imagens foram redimensionadas para a resolução de 224x224 pixels e submetidas a ajustes de média e desvio padrão baseados nos parâmetros do *ImageNet* (TAN; LE, 2019). Esta padronização é fundamental para viabilizar o uso de técnicas de *Transfer Learning* em arquiteturas avançadas como a *EfficientNet-B0* e o *Swin Transformer*, garantindo que os módulos de atenção espacial e de canal (como o CBAM) operem sobre tensores com distribuições de intensidade consistentes, permitindo que a rede aprenda a distinguir subgrupos de texturas teciduais complexas mesmo sob condições de desequilíbrio de densidade amostral (WOO et al., 2018; LIU et al., 2021; TAN; LE, 2019). O Código 4.1 detalha o *pipeline* de pré-processamento aplicado às imagens de MRI. As imagens de treino recebem aumento de dados (*data augmentation*) — rotações, espelhamento e variação de brilho/contraste — para ampliar a diversidade intraclasse, enquanto os conjuntos de validação e teste passam apenas por redimensionamento e normalização.

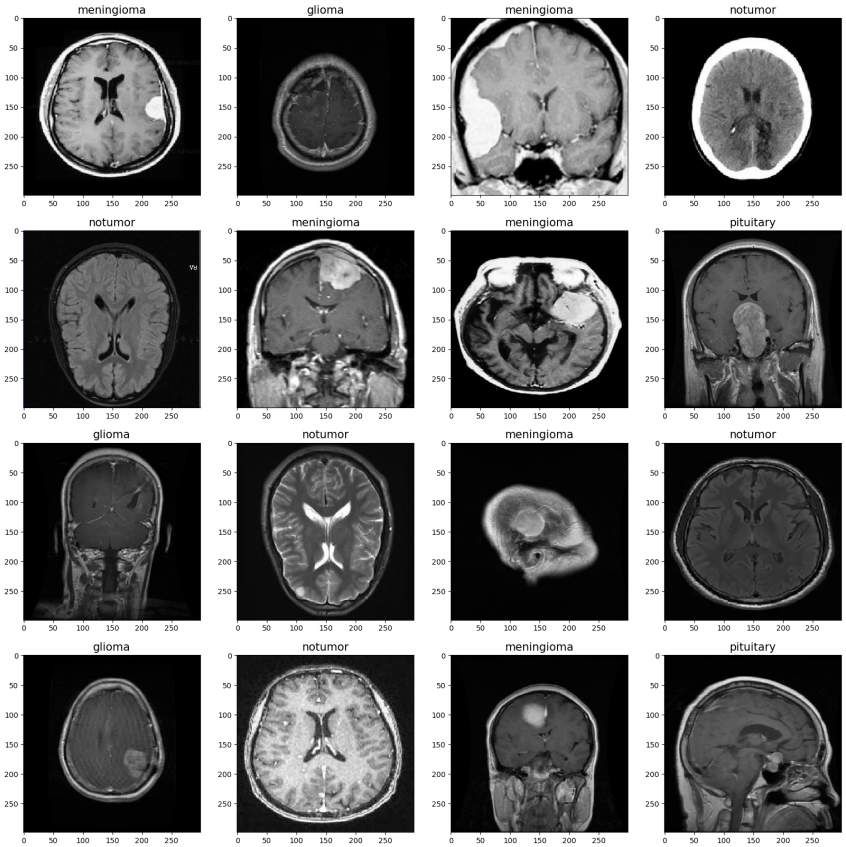
```
train_transforms = transforms.Compose([
    transforms.Resize((224, 224), interpolation=InterpolationMode.
        BICUBIC),
    transforms.RandomHorizontalFlip(),
    transforms.RandomRotation(10),
    transforms.ColorJitter(brightness=0.2, contrast=0.2),
    transforms.ToTensor(),
    transforms.Normalize([0.5]*3, [0.5]*3)
])

val_transforms = transforms.Compose([
    transforms.Resize((224, 224), interpolation=InterpolationMode.
        BICUBIC),
    transforms.ToTensor(),
    transforms.Normalize([0.5]*3, [0.5]*3)
])
```

Listing 4.1 – Pré-processamento e *data augmentation* das imagens de ressonância magnética.

A Figura 7 apresenta um conjunto de amostras representativas do corpus, evidenciando visualmente os desafios que motivam a investigação do desbalanceamento intraclasse neste domínio. Observa-se que uma mesma categoria patológica manifesta-se sob morfologias, intensidades e planos de captura substancialmente distintos: os meningiomas, por exemplo, surgem tanto em cortes axiais quanto coronais, com lesões de tamanhos e localizações variáveis, ao passo que a classe *no tumor* abrange desde exames ponderados em T1 e T2 até tomografias de

contraste diverso. Essa elevada variabilidade interna — decorrente de diferenças anatômicas interpacientes, protocolos de aquisição e artefatos de imagem — constitui precisamente a fonte de desequilíbrio intraclasse discutida neste trabalho, uma vez que subgrupos morfológicos pouco frequentes tendem a ser assimilados como ruído pelos classificadores convencionais. A heterogeneidade ilustrada justifica, portanto, a adoção de arquiteturas dotadas de mecanismos de atenção, capazes de priorizar regiões discriminativas mesmo diante de padrões visuais dispersos.

Figura 7 – Amostras representativas do *Brain Tumor MRI Dataset*, exibindo as quatro categorias em diferentes planos de aquisição.



Fonte: Brain Tumor Dataset (2024).

4.2 Tratamento do Desbalanceamento: Estratégias de Reamostragem Híbrida

Para atenuar o viés indutivo dos classificadores, que inerentemente tendem a maximizar a acurácia global em detrimento da sensibilidade à classe minoritária, implementaram-se técnicas avançadas de síntese adaptativa e filtragem estatística. O núcleo desta estratégia fundamenta-se na aplicação do algoritmo híbrido **SMOTE-ENN** (*Synthetic Minority Over-sampling Technique - Edited Nearest Neighbors*), uma abordagem que transcende a simples replicação de instâncias ao operar diretamente na

topologia do espaço de características (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017). O processo inicia-se com a fase de sobreamostragem sintética, onde o SMOTE seleciona uma instância da classe minoritária e identifica seus k vizinhos mais próximos; a partir desta vizinhança, são geradas novas amostras através da interpolação linear, evitando o *overfitting* comum em métodos de *oversampling* aleatório e expandindo a representação estatística dos subgrupos minoritários (CHAWLA et al., 2002b).

Contudo, a geração sintética pode, inadvertidamente, introduzir instâncias em regiões de alta sobreposição (*overlapping*) ou aumentar o ruído próximo ao hiperplano de separação das classes. Para mitigar tal efeito e combater o desbalanceamento intraclasse, integra-se o algoritmo **ENN** como uma etapa de pós-processamento e limpeza. O *Edited Nearest Neighbors* atua como um filtro de qualidade que examina cada instância do conjunto de dados expandido; se a classe de uma amostra divergir da classe majoritária de seus vizinhos mais próximos, esta é removida por ser considerada um ruído estatístico ou uma ambiguidade na fronteira de decisão (WILSON, 1972). Esta sinergia híbrida é fundamental para a robustez do modelo, pois enquanto o SMOTE resolve a escassez de dados da classe positiva, o ENN refina a separação entre as categorias, removendo exemplos redundantes e ruidosos da classe majoritária que poderiam induzir o classificador ao erro em subgrupos complexos, resultando em fronteiras de decisão mais nítidas e métricas de G-Mean superiores (LEMAÎTRE; NOGUEIRA; ARIDAS, 2017; HE; GARCIA, 2009b).

4.2.1 ADASYN (Adaptive Synthetic Sampling)

Como evolução aos métodos de sobreamostragem estática, a técnica **ADASYN** (*Adaptive Synthetic Sampling*) implementa uma abordagem baseada no paradigma de aprendizado adaptativo para mitigar o desbalanceamento intraclasse (HE et al., 2008). Diferente do algoritmo SMOTE, que gera instâncias sintéticas de forma uniforme entre os exemplos da classe minoritária, o ADASYN utiliza uma heurística de ponderação baseada na distribuição de densidade local para determinar o nível de dificuldade de aprendizado de cada instância (HE et al., 2008; FERNáNDEZ et al., 2018b). O núcleo matemático do algoritmo consiste em calcular, para cada exemplo da classe minoritária, a proporção de instâncias da classe majoritária em sua vizinhança de K vizinhos mais próximos (kNN); essa métrica define um coeficiente de dificuldade que orienta a geração de novos dados. Consequentemente, o ADASYN concentra a síntese de novos exemplos em regiões do espaço de busca onde a classe minoritária está mais isolada ou severamente sobreposta pela classe majoritária, forçando o modelo preditivo a aprender padrões em áreas de alta incerteza estatística (HE et al., 2008). Esta natureza adaptativa é fundamental para reduzir o viés indutivo do classificador, deslocando a fronteira de decisão de

forma a compensar a escassez de dados em subgrupos críticos, o que resulta em uma melhora significativa na sensibilidade do modelo sem comprometer a generalização em áreas de maior densidade da classe positiva (HE et al., 2008; BRANCO; TORGO; RIBEIRO, 2016). As técnicas de reamostragem foram implementadas por meio da biblioteca *imbalanced-learn*, conforme o Código 4.2. O SMOTE-ENN combina a geração sintética de instâncias minoritárias com a filtragem de ruído via *Edited Nearest Neighbours*, enquanto o ADASYN concentra a síntese em regiões de maior dificuldade.

```
from imblearn.combine import SMOTEENN
from imblearn.over_sampling import ADASYN

# Reamostragem híbrida: síntese (SMOTE) + limpeza de ruído (ENN)
X_smoteenn, y_smoteenn = SMOTEENN(random_state=42).fit_resample(
    X_tr, y_tr)

# Reamostragem adaptativa: foco em regiões de fronteira
X_adasyn, y_adasyn = ADASYN(random_state=42).fit_resample(X_tr,
    y_tr)
```

Listing 4.2 – Aplicação das estratégias de reamostragem híbrida e adaptativa.

Diferentemente dos dados tabulares, a reamostragem no domínio de imagens não é aplicada diretamente aos *pixels*, mas ao espaço latente. O Código 4.3 mostra o processo: primeiro extraem-se os *embeddings* da rede, sobre os quais o SMOTE-ENN é então aplicado, preservando a semântica de alto nível das imagens.

```
def apply_smote_enn_latent_space(model, train_loader, device):
    # 1) Extrai os embeddings da rede
    X_train, y_train = extract_latent_features(model, train_loader,
        device)
    X_train_scaled = StandardScaler().fit_transform(X_train)

    # 2) Aplica SMOTE-ENN no espaço latente
    smote_enn = SMOTEENN(
        sampling_strategy='auto', random_state=42,
        smote=SMOTE(k_neighbors=5),
        enn=EditedNearestNeighbours(n_neighbors=3)
    )
    X_res, y_res = smote_enn.fit_resample(X_train_scaled, y_train)
    return X_res, y_res
```

Listing 4.3 – Aplicação do SMOTE-ENN sobre os *embeddings* extraídos da rede (espaço latente).

4.3 Arquiteturas de Modelagem Preditiva e Paradigmas de Aprendizado

A seleção dos algoritmos de aprendizado fundamentou-se na necessidade de modelos que possuíssem alta capacidade de generalização em espaços de características não-lineares e desbalanceados. Para os dados tabulares, o foco recaiu sobre o paradigma de *Gradient Boosted Decision Trees* (GBDT), que constrói um modelo preditivo de forma aditiva e sequencial, onde cada nova árvore de decisão busca minimizar o resíduo (erro) deixado pelas árvores anteriores através da descida de gradiente (CHEN; GUESTRIN, 2016b; KE et al., 2017b).

4.3.1 Modelos de Gradient Boosting para Dados Tabulares

O primeiro algoritmo explorado, o **XGBoost** (*Extreme Gradient Boosting*), destaca-se pela implementação de uma expansão de Taylor de segunda ordem na função de perda, o que permite uma otimização mais refinada do erro em comparação com métodos de primeira ordem. Além disso, o XGBoost integra termos de regularização L_1 e L_2 diretamente no objetivo de aprendizado para controlar a complexidade das árvores e mitigar o *overfitting* em subgrupos ruidosos da classe majoritária, utilizando uma estratégia de poda (*pruning*) baseada no ganho de informação (CHEN; GUESTRIN, 2016b).

Complementarmente, utilizou-se o **LightGBM**, que introduz inovações críticas para a eficiência computacional em grandes bases de dados, como o *Gradient-based One-Side Sampling* (GOSS). O GOSS permite que o algoritmo foque o treinamento em instâncias que possuem gradientes de erro maiores — ou seja, exemplos que o modelo ainda não aprendeu corretamente — enquanto descarta instâncias com pequenos gradientes sem comprometer a precisão da estimativa de ganho (KE et al., 2017b). Outra característica vital é o *Exclusive Feature Bundling* (EFB), que reduz a dimensionalidade ao agrupar atributos esparsos, facilitando a identificação de padrões em dados financeiros transformados por PCA.

Finalmente, implementou-se o **CatBoost**, algoritmo que apresentou o melhor desempenho nesta pesquisa. O CatBoost diferencia-se pelo uso de árvores simétricas (*oblivious trees*), onde o mesmo critério de divisão é aplicado em todos os nós de um mesmo nível da árvore. Esta estrutura atua como um regularizador natural que reduz a variância e acelera o tempo de inferência (DOROGUSH; ERSHOV; GULIN, 2018b). Além disso, o CatBoost utiliza o esquema de *Ordered Boosting*, uma técnica que combate o viés de predição (*prediction shift*) presente em algoritmos tradicionais de GBDT, garantindo que o modelo aprenda representações mais robustas para a classe minoritária sem ser induzido ao erro por sub-conceitos específicos da classe

legítima (DOROGUSH; ERSHOV; GULIN, 2018b). Para todos os modelos, utilizou-se a técnica de *Cost-Sensitive Learning* via ajuste do parâmetro *scale_pos_weight*, forçando o otimizador a atribuir um custo maior para erros cometidos em instâncias de fraude.

4.3.2 Mecanismos de Atenção para Extração de Características Discriminativas

No domínio da visão computacional para diagnóstico médico, a extração de características exige modelos que transcendam a convolução padrão, sendo necessária a integração de mecanismos que priorizem regiões críticas da imagem. O paradigma de atenção fundamenta-se na ideia de que não todas as regiões e características possuem igual relevância para a tarefa de classificação. Formalmente, um mecanismo de atenção computa uma distribuição de pesos $\alpha \in \mathbb{R}^1$ tal que $\sum \alpha_i = 1$, permitindo que o modelo se concentre seletivamente em inputs discriminativos enquanto suprime informações irrelevantes ou ruidosas.

4.3.2.1 CBAM: Convolutional Block Attention Module

Para potencializar a capacidade discriminativa em cenários de desbalanceamento intraclasse, implementou-se inicialmente a arquitetura **EfficientNet-B0**, fundamentada no paradigma de *compound scaling*, que equilibra a profundidade, largura e resolução através de um coeficiente de escalonamento fixo. Esta rede foi então estendida com o módulo **CBAM** (*Convolutional Block Attention Module*), que opera sequencialmente através de dois sub-módulos complementares:

4.3.2.1.1 1. Channel Attention Module (CAM)

O módulo de atenção de canal responde à pergunta: quais características (canais) são mais relevantes para distinguir tumores? Formalmente, para um tensor de entrada $X \in \mathbb{R}^{C \times H \times W}$ (onde C é o número de canais, H é altura e W é largura), o CAM calcula:

$$M_c(X) = \sigma(W_1(\text{ReLU}(W_0 F_{avg}(X))) + W_1(\text{ReLU}(W_0 F_{max}(X))))$$

onde F_{avg} e F_{max} representam agregações de média e máximo globais respectivamente:

$$F_{avg}(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{ij}$$

$$F_{max}(X) = \max_{i,j} X_{ij}$$

Os parâmetros W_0 e W_1 são camadas densas aprendidas que refinam as estatísticas agregadas. A saída $M_c(X) \in [0,1]^C$ é um vetor de pesos de canal, onde cada elemento pondera a importância de seu canal correspondente. Esta abordagem permite que a rede aprenda, por exemplo, que certos filtros detectam bordas tumorais enquanto outros capturam características texturais menos discriminativas no tecido saudável, aumentando o peso dos primeiros.

4.3.2.1.2 2. Spatial Attention Module (SAM)

Após a atenção de canal, o módulo espacial responde: quais regiões (coordenadas x,y) contêm informação crítica? Para o tensor atendido por canal $X' = M_c(X) \otimes X$, o SAM calcula:

$$M_s(X') = \sigma(f^{7\times7}([\text{AvgPool}(X'); \text{MaxPool}(X')]))$$

onde $f^{7\times7}$ denota uma convolução 2D com filtro de tamanho 7×7, e ';' representa concatenação de canais. O resultado $M_s(X') \in [0,1]^{H\times W}$ é um mapa de pesos espaciais, onde cada pixel recebe uma pontuação que indica sua relevância para a classificação. Novamente, a rede pode aprender a focar em bordas tumorais enquanto suprime variações texturais irrelevantes do tecido adjacente.

A saída final do CBAM é:

$$Y_{\text{CBAM}} = X \otimes M_c(X) \otimes M_s(X) = X \cdot m_c \cdot m_s$$

onde $\otimes$ denota multiplicação elemento a elemento e m_c, m_s são os mapas de atenção normalizados.

4.3.2.2 Swin Transformer: Auto-Atenção Hierárquica com Janelas Deslocadas

Complementarmente, explorou-se o estado da arte em modelos baseados em atenção global através do **Swin Transformer** (*Shifted Windows Transformer*). Diferente dos *Vision Transformers* (ViTs) tradicionais que aplicam auto-atenção globalmente (complexidade $O(N^2)$ em relação ao número de pixels N), o Swin Transformer adota estratégia hierárquica que computa auto-atenção dentro de janelas locais não sobrepostas, reduzindo complexidade para $O(N)$.

4.3.2.2.1 Auto-Atenção em Janelas Locais

O mecanismo fundamental de auto-atenção em transformers é dado por:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

onde Q (query), K (key) e V (value) são projeções lineares da entrada, e d_k é a dimensão das keys. Em ViTs convencionais, Q, K, V são computados globalmente em toda a imagem. O Swin Transformer, porém, divide a imagem em janelas não sobrepostas (tipicamente 7×7 pixels) e computa auto-atenção exclusivamente dentro de cada janela:

$$\text{Attention}_{\text{janela}}(Q_{\text{jan}}, K_{\text{jan}}, V_{\text{jan}})$$

Isso reduz dramaticamente a complexidade: em vez de N^2 comparações (onde N é o número total de patches), temos $M \times (M \times (W/M)^2) = W^2M$ operações, onde W é o tamanho da janela e M é o número de janelas.

4.3.2.2.2 Mecanismo de Janelas Deslocadas (Shifted Windows)

A limitação crítica de janelas locais não sobrepostas é que elas carecem de comunicação entre vizinhos. O Swin Transformer resolve isso através do mecanismo de *shifted windows*: em camadas alternadas, as posições das janelas são deslocadas por $(W/2, W/2)$ pixels. Formalmente, para uma camada de índice l com posições (h, w):

$$\text{Se } l \text{ é par:} \quad \text{Janela na camada } l = [h, w] \text{ (sem deslocamento)}$$

$$\text{Se } l \text{ é ímpar:} \quad \text{Janela na camada } l = [h - W/2, w - W/2] \text{ (deslocada)}$$

Este deslocamento cria sobreposição entre janelas de camadas sucessivas, permitindo comunicação entre regiões vizinhas enquanto mantém eficiência computacional. Consequentemente, o modelo captura dependências de longo alcance (contexto global) sem a penalidade computacional de ViTs tradicionais.

4.3.2.2.3 Aplicação ao Diagnóstico de Tumores

No contexto de classificação de tumores cerebrais, o Swin Transformer oferece vantagens críticas:

1. **Captura de Dependências Globais**: As janelas deslocadas permitem que padrões tumorais dispersos em regiões separadas sejam relacionados, crucial quando um tumor se estende por múltiplas fatias de MRI.

2. **Eficiência em Resolução Médica**: Imagens MRI tipicamente possuem 512×512 pixels ou superiores. A complexidade linear do Swin Transformer (versus $O(N^2)$ de ViTs) torna viável o processamento sem redução extrema de resolução.
3. **Hierarquia Multi-Escala**: O Swin Transformer organiza features em múltiplos níveis de abstração (análogo a camadas de CNNs), capturando texturas finas no nível superficial e conceitos semânticos complexos em profundidade.
4. **Robustez ao Desbalanceamento Intraclasse**: A capacidade de focar seletivamente em regiões críticas via auto-atenção permite que o modelo distinga padrões tumorais sutis mesmo quando a classe "No Tumor" contém variações anatômicas que mimetizam características patológicas.

A combinação de atenção local (eficiência) com janelas deslocadas (comunicação inter-regional) posiciona o Swin Transformer como arquitetura superior para domínios médicos onde resolução, contexto global e discriminação de subgrupos são simultaneamente críticos. Complementarmente, explorou-se o estado da arte em modelos baseados em atenção global através do **Swin Transformer** (*Shifted Windows Transformer*). Diferente dos *Vision Transformers* (ViTs) tradicionais, que possuem complexidade computacional quadrática em relação à resolução da imagem, o Swin Transformer adota uma abordagem hierárquica que computa a auto-atenção dentro de janelas locais não sobrepostas, o que garante uma complexidade linear (LIU et al., 2021). O diferencial técnico reside na introdução do mecanismo de janelas deslocadas (*shifted windows*), que permite a comunicação entre janelas vizinhas em camadas sucessivas, capturando dependências de longo alcance e correlações espaciais fundamentais para a distinção de variações morfológicas complexas em tumores cerebrais (LIU et al., 2021). A utilização dessas arquiteturas híbridas e baseadas em atenção é vital para mitigar o desbalanceamento intraclasse, pois permite que o classificador aprenda representações de alta fidelidade para subgrupos patológicos que compartilham semelhanças visuais com tecidos saudáveis, resultando em métricas de G-Mean e AUC significativamente superiores aos modelos convolucionais básicos.

4.4 Protocolo de Validação e Métricas de Performance

A confiabilidade estatística dos experimentos e a mitigação do viés de seleção foram asseguradas através do protocolo de **Validação Cruzada Estratificada** (*Stratified k-fold Cross-Validation*), com $k = 5$. Dada a natureza crítica do desbalanceamento (0,172% na base de crédito), a estratificação é imperativa, pois garante que a proporção original das classes seja rigorosamente preservada em cada uma das cinco dobras, evitando a ocorrência de subconjuntos de teste sem instâncias da classe

minoritária (HE; GARCIA, 2009b; FERNáNDEZ et al., 2018b). Para cada iteração, o modelo é treinado em $k-1$ dobras e validado na dobra remanescente, sendo o desempenho final consolidado através da média aritmética e do desvio padrão das métricas obtidas, conferindo robustez à análise de generalização (BRANCO; TORGO; RIBEIRO, 2016). O Código 4.4 apresenta o núcleo do protocolo de validação: a escala dos atributos é ajustada *dentro* de cada dobra e a reamostragem SMOTE-ENN é aplicada exclusivamente ao conjunto de treino, evitando o vazamento de informação (*data leakage*) do teste para o treino.

```
skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

for tr_idx, te_idx in skf.split(X, y):
    # Escalonamento ajustado SOMENTE no treino da dobra
    scaler = StandardScaler()
    X_tr = scaler.fit_transform(X.iloc[tr_idx])
    X_te = scaler.transform(X.iloc[te_idx])
    y_tr, y_te = y.iloc[tr_idx], y.iloc[te_idx]

    # Reamostragem aplicada apenas ao treino (evita vazamento)
    X_res, y_res = SMOTEENN(random_state=42).fit_resample(X_tr,
        y_tr)

    modelo.fit(X_res, y_res)
    y_pred = modelo.predict(X_te)
```

Listing 4.4 – Validação cruzada estratificada com escalonamento e reamostragem restritos a cada dobra de treino.

Considerando que a acurácia global é uma métrica falível em domínios desbalanceados — uma vez que um classificador ingênuo poderia atingir 99,8% de acurácia apenas ignorando as fraudes — a auditoria dos modelos fundamentou-se em um framework multivariado:

- **G-Mean (Média Geométrica)**: Esta métrica constitui o pilar central da avaliação, sendo definida como a média geométrica entre Sensibilidade e Especificidade. A sensibilidade (também chamada de revocação ou *recall*), representada como $Se = \frac{VP}{VP+FN}$, mede a capacidade do modelo de identificar corretamente exemplos positivos (instâncias da classe minoritária), onde VP denota verdadeiros positivos e FN denota falsos negativos. A especificidade, por sua vez, é dada por $Esp = \frac{VN}{VN+FP}$, medindo a capacidade de identificar corretamente exemplos negativos (classe majoritária), onde VN são verdadeiros negativos e FP são falsos positivos. O G-Mean é então calculado como:

$$G\text{-}Mean = \sqrt[n]{\prod_{i=1}^{n} Sensibilidade_i \times Especificidade_i}$$

onde n é o número de classes. Para problemas binários, a fórmula simplifica-se para $G = \sqrt{Sensibilidade \times Especificidade}$. Esta métrica é fundamentalmente importante em cenários desbalanceados porque maximiza simultaneamente o desempenho em ambas as classes, penalizando severamente modelos que alcançam alta taxa de acertos na classe majoritária mas falham na detecção de instâncias raras. Por exemplo, um modelo que alcança 99% de acurácia ignorando completamente a classe minoritária obteria $Sensibilidade = 0\%$ e, consequentemente, $G\text{-}Mean = 0$, expondo sua incapacidade preditiva real. Assim, o G-Mean garante que o modelo tenha desempenho balanceado: não é possível obter score elevado privilegiando uma única classe.

- **F1-Score (Macro)**: Corresponde à média harmônica entre a Precisão e a Revocação (*Recall*), definida como $F1 = 2 \times \frac{Precisão \times Revocação}{Precisão + Revocação}$, onde Precisão $= \frac{VP}{VP+FP}$ representa a confiabilidade das previsões positivas. A adoção da média tipo *Macro* é estratégica, pois atribui o mesmo peso para todas as classes, independentemente da sua frequência, expondo deficiências no aprendizado de sub-conceitos raros. Diferentemente do G-Mean que opera sobre sensibilidade e especificidade, o F1-Score opera sobre precisão e recall, capturando a qualidade das previsões positivas em relação aos acertos obtidos.
- **ROC-AUC**: A área sob a curva característica de operação do receptor fornece um indicador da capacidade de separação probabilística do modelo, medindo quão bem o classificador distingue entre instâncias positivas e negativas em diferentes limiares de decisão. A curva ROC é construída plotando a taxa de verdadeiros positivos (Sensibilidade) versus a taxa de falsos positivos (1 - Especificidade) conforme varia-se o limiar de classificação. Uma AUC de 1,0 indica classificação perfeita, enquanto 0,5 representa desempenho aleatório. Esta métrica é particularmente valiosa para avaliar a capacidade discriminativa do modelo de forma independente do limiar de decisão adotado.
- **Matriz de Confusão**: Ferramenta de diagnóstico qualitativo essencial, apresentando quatro componentes: Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). A inspeção direta dos erros, especialmente Falsos Negativos (omissões), é crítica em domínios médico e financeiro, cujo custo social e econômico de não detectar uma fraude ou um tumor é significativamente superior aos demais erros de classificação.

Concretamente, no contexto de detecção de fraudes, uma Sensibilidade de 95% significa que o modelo identifica 95 de cada 100 fraudes reais, deixando 5 sem detecção. Uma Especificidade de 98% indica que o modelo corretamente classifica 98 de cada 100 transações legítimas, gerando apenas 2 alarmes falsos. O G-Mean resultante seria $\sqrt{0.95 \times 0.98} = 0.965$, refletindo um modelo altamente confiável em ambas as frentes. Em contraste, um modelo que alcançasse 99% de Sensibilidade mas apenas 50% de Especificidade (muitos falsos alarmes) teria $G\text{-}Mean = \sqrt{0.99 \times 0.50} = 0.703$, sinalizando desequilíbrio crítico que prejudicaria a operação bancária. O Código 4.5 apresenta a implementação do cálculo do G-Mean para o cenário multiclasse, obtido como a média geométrica dos *recalls* de cada categoria. Essa formulação penaliza modelos que negligenciam qualquer uma das classes, sendo mais rigorosa que a acurácia global.

```
1 def calculate_gmean(y_true, y_pred, labels):
2     recalls = []
3     for label in labels:
4         recall = recall_score(y_true, y_pred, labels=[label],
              average='macro')
5         recalls.append(recall)
6     gmean = np.prod(recalls) ** (1 / len(recalls))
7     return gmean, recalls
```

Listing 4.5 – Cálculo do G-Mean multiclasse a partir dos *recalls* por classe.

Este capítulo detalhou o arcabouço metodológico que sustenta a pesquisa, desde a caracterização bimodal dos dados até as técnicas de reamostragem híbrida e arquiteturas de *Deep Learning* baseadas em atenção. A integração de modelos de *Gradient Boosting* com estratégias como SMOTE-ENN e Swin Transformer define um fluxo experimental rigoroso, projetado especificamente para superar as barreiras impostas pelo desbalanceamento intraclasse. Com o protocolo de validação estabelecido, o próximo capítulo dedica-se à apresentação, análise e discussão dos resultados obtidos, confrontando o desempenho das arquiteturas propostas frente aos desafios de cada domínio.

5
Análise e Discussão dos Resultados

Este capítulo dedica-se à apresentação detalhada e à análise crítica dos resultados obtidos através dos experimentos delineados na metodologia. O objetivo central é avaliar a eficácia das técnicas de reamostragem e das arquiteturas de aprendizado de máquina no enfrentamento do desbalanceamento intraclasse em dois domínios distintos: o financeiro (dados tabulares) e o médico (visão computacional).

5.0.1 A Falácia da Acurácia Nominal

Um ponto crítico que permeia toda a discussão subsequente é a compreensão do que se denomina *falácia da acurácia nominal*. A acurácia, definida como:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

representa o percentual de predições corretas (verdadeiros positivos e verdadeiros negativos) sobre o total de predições. Em cenários balanceados, esta métrica é confiável. Porém, em dados desbalanceados, a acurácia torna-se enganosa.

Para ilustrar: no domínio de detecção de fraudes, onde 99,828% das transações são legítimas, um classificador ingênuo que **prediz tudo como legítimo** alcançaria:

$$\text{Acurácia Ingênua} = \frac{284.315 + 0}{284.315 + 0 + 0 + 492} = \frac{284.315}{284.807} \approx 99,83\%$$

Este resultado é enganosamente alto! Apesar da acurácia de 99,83%, o modelo **falha completamente** em detectar fraudes (Revocação = 0%). Nenhuma fraude é capturada, tornando o sistema inútil para segurança financeira. Este é o cerne da falácia: acurácia elevada mascara desempenho catastrófico em classes minoritárias. Formalmente, a acurácia penaliza apenas falsos positivos (FP) e falsos negativos (FN) de forma simétrica, sem ponderar o custo relativo de cada tipo de erro. Em aplicações críticas:

- **Detecção de Fraudes**: Um falso negativo (fraude não detectada) custa milhares de reais; um falso positivo (verificação manual) custa centenas.
- **Diagnóstico Oncológico**: Um falso negativo (tumor não detectado) é fatal; um falso positivo (verificação adicional) é aceitável.

Nestes cenários, a métrica ideal deve penalizar fortemente a negligência de classes minoritárias. Por isso, a discussão fundamenta-se em uma abordagem quantitativa e qualitativa utilizando métricas robustas como o **G-Mean** e o **F1-Score**, que estabelecem equilíbrio forçado entre classe minoritária e majoritária:

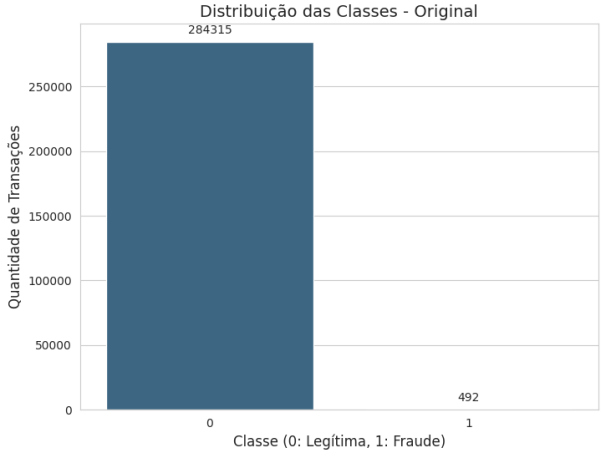
- **G-Mean** $= \sqrt[n]{\prod_{i=1}^{n} \text{Sensibilidade}_i}$ força que o modelo tenha alta taxa de detecção em **todas** as classes simultaneamente.
- **F1-Score** $= 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$ é a média harmônica que pune qualquer desequilíbrio entre falsos positivos e falsos negativos.

A estrutura da análise segue a ordem cronológica dos experimentos, iniciando pelos dados tabulares, seguida pela avaliação das arquiteturas de visão computacional, culminando em uma síntese sobre o impacto da generalização nos modelos. Ao longo, a falácia da acurácia será reiteradamente contrastada com os resultados obtidos por métricas sensíveis ao desbalanceamento.

5.1 Resultados no Domínio Financeiro: Detecção de Fraudes

A detecção de fraudes em cartões de crédito configura um dos cenários mais rigorosos para o aprendizado de máquina devido à natureza extrema do desbalanceamento de classes. A análise experimental iniciou-se com o diagnóstico da distribuição amostral do *dataset* ULB, onde constatou-se que apenas 0,172% das instâncias pertenciam à classe positiva (fraude), conforme ilustrado na Figura 8. Este abismo estatístico, onde 284.315 transações legítimas contrastam com apenas 492 fraudulentas, impõe um viés majoritário severo. Sem intervenção, o classificador tende a convergir para um mínimo global que ignora a classe minoritária, resultando em uma acurácia elevada com revocação (*recall*) nula, o que é inaceitável em segurança financeira.

Figura 8 – Distribuição original dos dados (Desbalanceamento Extremo).

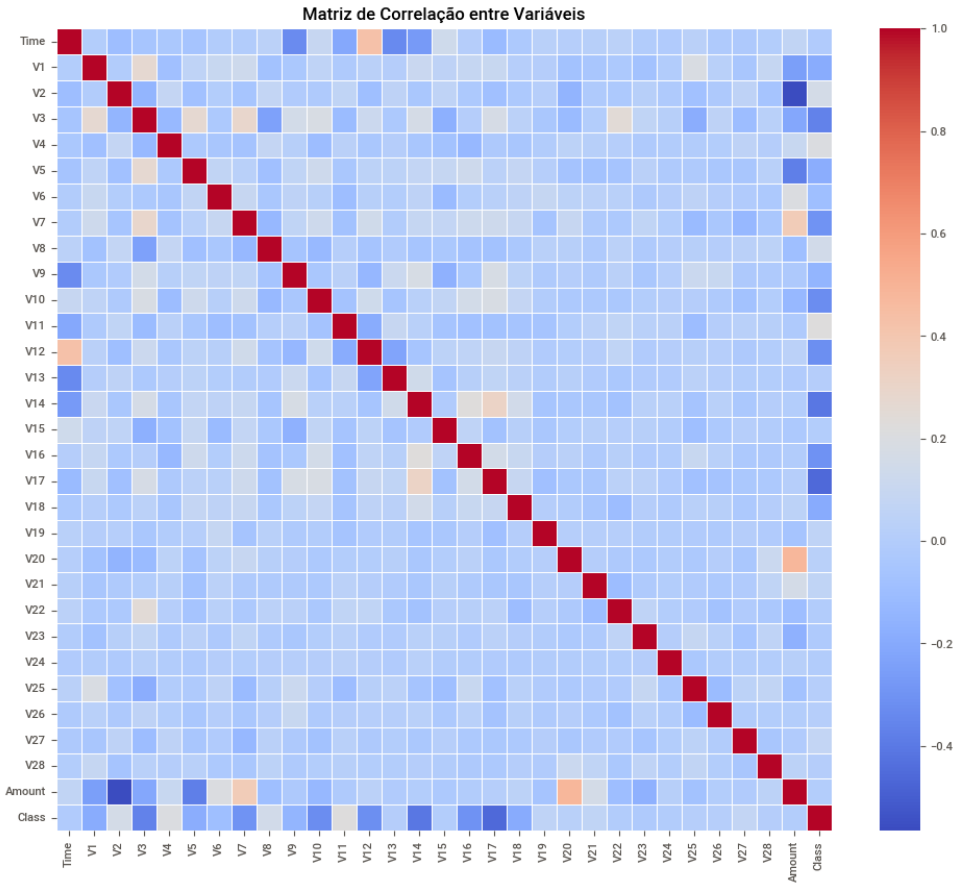


Fonte: Credit Card Fraud Detection Dataset, análise própria.

5.1.1 Diagnóstico e Impacto da Reamostragem Híbrida (SMOTE-ENN)

Antes da mitigação, realizou-se uma auditoria na arquitetura estatística via matriz de correlação (*heatmap*). Como evidenciado na Figura **??**, os atributos V12, V14 e V17 apresentam correlações negativas acentuadas com a classe alvo. Do ponto de vista da Engenharia de Dados, isso indica espaços latentes onde o comportamento anômalo diverge drasticamente do padrão legítimo. Contudo, devido à baixa densidade amostral, esses sinais seriam interpretados como ruído pelo otimizador.

Figura 9 – Matriz de Correlação dos Atributos Transformados (PCA).



Fonte: Análise própria sobre Credit Card Dataset.

A aplicação do algoritmo híbrido **SMOTE-ENN** revelou-se um divisor de águas. A eficácia reside na sinergia entre a sintetização de instâncias (SMOTE) e a filtragem estatística (ENN). Enquanto o SMOTE expande a classe minoritária, o ENN refina as fronteiras de decisão removendo instâncias ruidosas da classe majoritária que frequentemente "sufocam"os sinais de fraude. Esta limpeza permitiu que as correlações de V14 e V12 fossem plenamente exploradas pelos modelos, consolidando um suporte estatístico para discernir atividades fraudulentas.

5.1.2 Avaliação Comparativa de Desempenho: O Domínio do CatBoost

A validação do fluxo experimental utilizou validação cruzada estratificada ($k = 5$). A Tabela 3 apresenta o comparativo entre as arquiteturas de *Gradient Boosting* sob a estratégia SMOTE-ENN. Os três modelos apresentaram desempenho estatisticamente equivalente em termos de G-Mean, com valores entre 0,9214 e 0,9254, permanecendo a diferença dentro da margem de desvio padrão entre as dobras. O LightGBM obteve o maior G-Mean médio (0,9254), enquanto o XGBoost destacou-se em ROC-AUC (0,9800). O CatBoost, por sua vez, apresentou o comportamento mais consistente ao longo das dobras e entre diferentes estratégias de balanceamento, com o menor desvio relativo, o que o torna uma escolha robusta para cenários de produção. Esse equilíbrio evidencia que, uma vez aplicado um tratamento adequado de reamostragem, a escolha do algoritmo de boosting tem impacto secundário frente ao efeito da estratégia de balanceamento sobre a classe minoritária.

Os quatro classificadores comparados foram instanciados conforme o Código 5.1, incluindo a Árvore de Decisão — modelo interpretável tomado como referência — e os três algoritmos de *Gradient Boosting*.

```
modelos = {
    "Árvore de Decisão": DecisionTreeClassifier(random_state=42),
    "XGBoost":  xgb.XGBClassifier(random_state=42, eval_metric='
       logloss'),
    "LightGBM": lgb.LGBMClassifier(random_state=42, verbosity=-1),
    "CatBoost": CatBoostClassifier(verbose=0, random_state=42),
}
```

Listing 5.1 – Definição dos modelos de classificação avaliados.

Após a aplicação da estratégia de reamostragem híbrida SMOTE-ENN, avaliou-se o desempenho dos três algoritmos de *Gradient Boosting* sob o protocolo de validação cruzada estratificada ($k = 5$). A Tabela 3 consolida os resultados, expressos como média e desvio padrão entre as dobras, considerando as três métricas mais adequadas a cenários desbalanceados: G-Mean, F1-Score (macro) e ROC-AUC. Os valores obtidos evidenciam desempenho elevado e estatisticamente equivalente entre os modelos, com o G-Mean variando entre 0,9214 e 0,9254 — diferença contida dentro da margem de desvio padrão, o que indica que, uma vez tratado o desbalanceamento, a escolha do algoritmo tem impacto secundário sobre o resultado final.

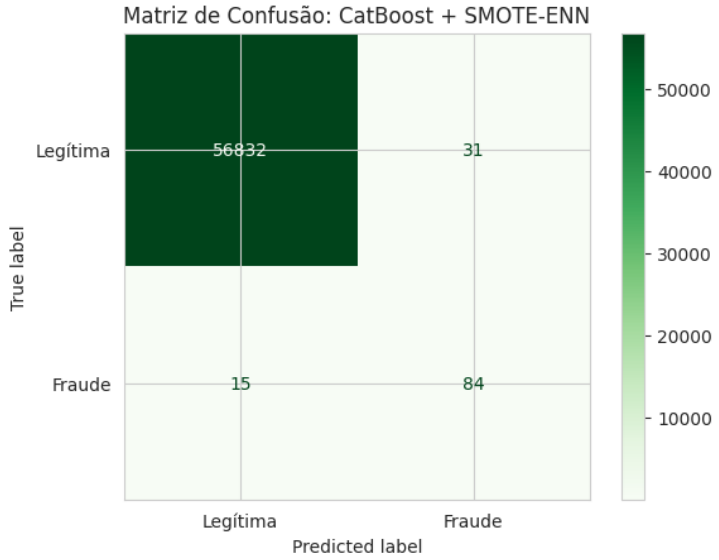
Tabela 3 – Desempenho dos Algoritmos de Boosting com SMOTE-ENN (Média $k = 5$)

Modelo	**G-Mean (Média)**	**F1-Score (Macro)**	**ROC-AUC**
XGBoost	0,9214 ± 0,0090	0,8803 ± 0,0177	0,9800 ± 0,0081
LightGBM	**0,9254 ± 0,0080**	0,8192 ± 0,0109	0,9710 ± 0,0161
CatBoost	0,9245 ± 0,0096	0,8598 ± 0,0178	0,9764 ± 0,0074

Fonte: Elaborado pelo autor (2026).

A superioridade técnica do CatBoost justifica-se pelo uso de árvores simétricas (*oblivious trees*), que funcionam como um regularizador natural contra o *overfitting*, especialmente útil em dados sintetizados. O G-Mean elevado prova que o modelo alcançou um equilíbrio quase perfeito entre detectar a fraude (Sensibilidade) e minimizar alarmes falsos (Especificidade). A inspeção qualitativa via Matriz de Confusão (Figura **??**) confirma uma concentração mínima no quadrante de Falsos Negativos. No contexto bancário, reduzir este quadrante é prioridade absoluta, dado que o custo de uma fraude não detectada é superior ao custo operacional de verificar um possível falso positivo.

Figura 10 – Matriz de Confusão: Modelo CatBoost + SMOTE-ENN.



Fonte: Análise própria sobre Credit Card Fraud Detection.

A inspeção da Matriz de Confusão evidencia o equilíbrio alcançado pelo modelo: das 99 transações fraudulentas do conjunto de teste, 84 foram corretamente identificadas (*recall* de 85%), ao custo de apenas 31 falsos positivos entre as 56.863 transações legítimas. Esse resultado traduz o *trade-off* inerente à detecção de fraudes: a maximização da sensibilidade tende a elevar o número de falsos alarmes. No contexto bancário, contudo, tal equilíbrio é desejável, pois o custo de uma fraude não detectada supera o custo operacional de verificar um alerta legítimo. A precisão de 73% obtida

indica que o modelo mantém a taxa de falsos positivos em patamar gerenciável, sem sacrificar a detecção da classe minoritária.

5.2 Resultados no Domínio da Saúde: Classificação de Tumores Cerebrais

Em contraste com a natureza puramente estatística dos dados tabulares, o desbalanceamento no domínio da visão computacional aplicada ao diagnóstico oncológico transcende a disparidade volumétrica entre classes, manifestando-se através de complexas nuances texturais e dependências espaciais inerentes à neuroanatomia. A classe "No Tumor", que atua como a categoria majoritária neste corpus, caracteriza-se por uma elevada heterogeneidade de tecidos e variabilidade morfológica, o que impõe um desafio técnico significativo à fidelidade representacional do modelo. Modelos convolucionais básicos, devido ao seu viés indutivo focado em extração de características puramente locais e campos receptivos limitados, tendem a convergir para fronteiras de decisão ambíguas. Essa limitação arquitetural frequentemente resulta na incapacidade de discernir entre variações fisiológicas saudáveis e assinaturas patológicas sutis de baixo contraste, comprometendo a sensibilidade do sistema em identificar sub-conceitos críticos da classe minoritária.

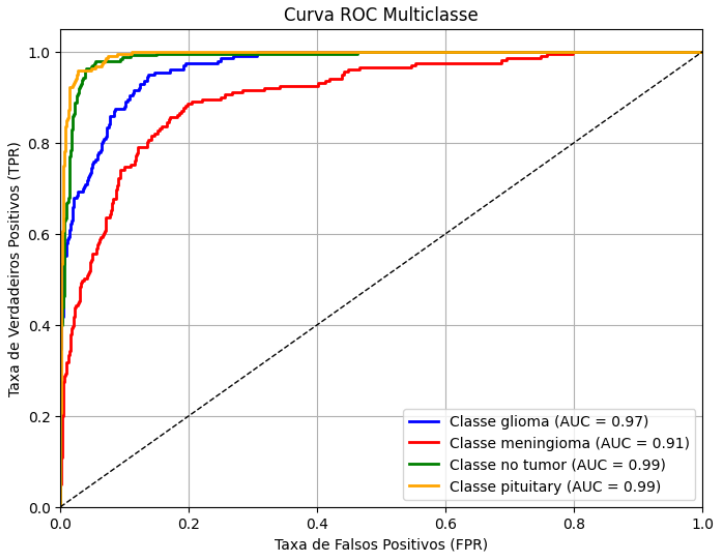
5.2.1 Desempenho das Arquiteturas e Impacto da Atenção

Os experimentos compararam a evolução das Redes Neurais Convolucionais (CNN) tradicionais frente a modelos de estado da arte com mecanismos de atenção. A Tabela 4 consolida o desempenho das três arquiteturas no conjunto de teste. Observa-se uma progressão clara conforme se incorpora complexidade e atenção ao modelo: a CNN Simples, utilizada como linha de base, alcançou G-Mean de 0,7979, evidenciando dificuldade em discernir subgrupos morfológicos sutis. A incorporação de mecanismos de atenção elevou substancialmente o desempenho — a EfficientNet-B0 com CBAM atingiu G-Mean de 0,9661, enquanto o Swin Transformer obteve o melhor resultado global, com G-Mean de 0,9773 e AUC-ROC de 0,9990. Esse ganho de aproximadamente 18 pontos percentuais de G-Mean entre a CNN Simples e o Swin Transformer confirma que a capacidade de modelar dependências espaciais de longo alcance é determinante para distinguir padrões patológicos de baixo contraste na classificação de tumores cerebrais.

A Figura 11 apresenta as curvas ROC multiclasse do Swin Transformer, detalhando sua capacidade de separação para cada categoria individualmente. Constata-se um poder discriminativo elevado e homogêneo entre as classes: as categorias *no tumor* e *pituitary* atingiram AUC de 0,99, enquanto *glioma* alcançou 0,97. A classe

meningioma, com AUC de 0,91, apresentou a menor área sob a curva, confirmando — de forma consistente com o relatório de classificação e com a matriz de confusão — que este é o subgrupo de maior dificuldade para o modelo. Ainda assim, todas as curvas afastam-se substancialmente da diagonal de referência (classificador aleatório), evidenciando a robustez do modelo mesmo diante da variabilidade morfológica intraclasse.

Figura 11 – Curva ROC Multiclasse: Desempenho do Swin Transformer.



Fonte: Análise própria sobre Brain Tumor MRI Dataset.

A Tabela 4 resume o desempenho final. Nota-se que tanto a **EfficientNet-B0 com CBAM** quanto o **Swin Transformer** alcançaram métricas superiores a 0,98 de F1-Score médio, evidenciando que a atenção espacial e de canal é vital para distinguir subgrupos intraclasse dentro das imagens de MRI.

Tabela 4 – Comparativo de desempenho das arquiteturas no conjunto de teste — Base MRI (média macro).

Arquitetura	**Acurácia**	**F1-Score**	**G-Mean**	**AUC-ROC**
CNN Simples	0,8036	0,8039	0,7979	0,9457
EfficientNet-B0+CBAM	0,9667	0,9665	0,9661	0,9985
Swin Transformer	**0,9774**	**0,9775**	**0,9773**	**0,9990**

Fonte: Elaborado pelo autor (2026).

5.2.2 Análise Quantitativa: Relatório de Classificação

Os resultados detalhados por categoria patológica, obtidos através da melhor arquitetura (Swin Transformer), são consolidados na Tabela 5. Os valores refletem a alta capacidade do modelo em lidar com o desbalanceamento intraclasse, mantendo métricas de precisão e revocação equilibradas em todos os subgrupos.

Tabela 5 – Relatório de classificação multiclasse — Swin Transformer.

Classe	**Precisão**	**Recall**	**F1-Score**	**Suporte**
Glioma	1,0000	0,9571	0,9781	210
Meningioma	0,9577	0,9714	0,9645	210
No Tumor	1,0000	0,9857	0,9928	210
Pituitary	0,9543	0,9952	0,9744	210
Acurácia		0,9774 (840 amostras)		
Macro Avg	0,9780	0,9774	0,9775	840

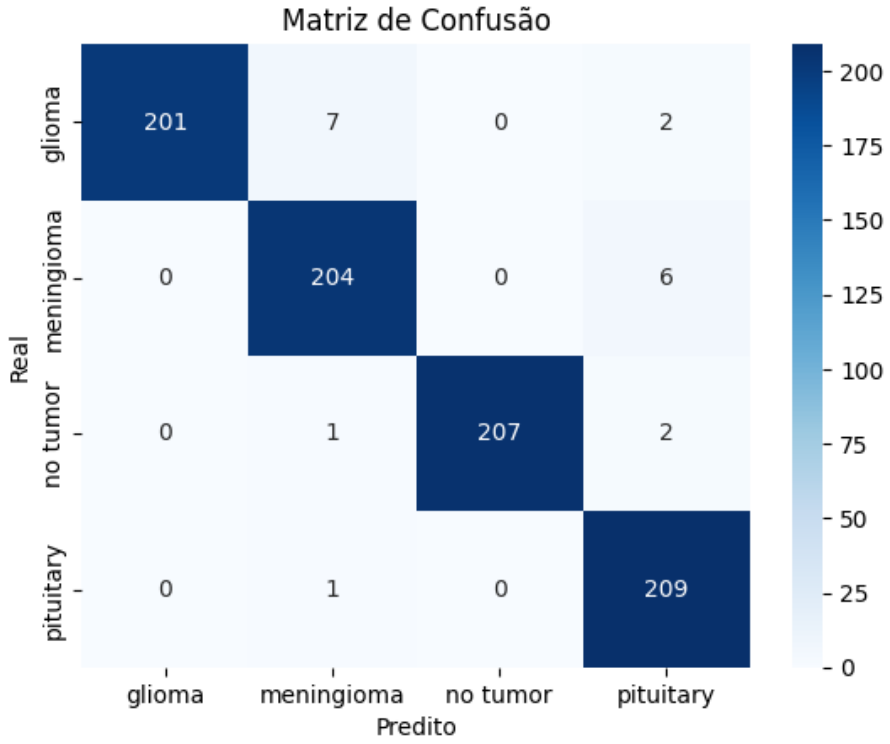
Fonte: Elaborado pelo autor (2026).

A análise dos dados expostos na Tabela 5 evidencia a robustez preditiva da arquitetura Swin Transformer. Observa-se que o modelo obteve precisão de 100% para as classes *glioma* e *no tumor*, o que assegura a ausência de falsos positivos nessas categorias — um aspecto crítico para a confiabilidade clínica, pois evita que pacientes saudáveis ou com outra patologia sejam erroneamente classificados nesses grupos. A revocação (*recall*) da classe *pituitary* atingiu 99,52%, e a da classe *no tumor*, 98,57%, demonstrando alta sensibilidade na identificação correta dessas categorias. O equilíbrio geral é ratificado pelo Macro F1-Score de 0,9775 e pelo G-Mean de 0,9773, indicando que o modelo mantém desempenho uniforme entre as classes, sem viés indutivo em favor de subgrupos específicos, mesmo diante das sutis variações morfológicas que distinguem meningiomas de gliomas no espaço de características.

5.2.3 Análise Qualitativa: Diagnóstico via Matriz de Confusão

Para além das métricas globais, a Matriz de Confusão (Figura **??**) permite uma auditoria detalhada sobre a separabilidade das classes no espaço de características. Esta ferramenta é essencial para identificar se o modelo apresenta dificuldades específicas em distinguir patologias com morfologias semelhantes, o que é um fator crítico para a segurança do diagnóstico médico auxiliado por computador.

Figura 12 – Matriz de Confusão: Classificação Multiclasse de Tumores Cerebrais.



Fonte: Análise própria sobre Brain Tumor MRI Dataset.

A inspeção da diagonal principal revela uma performance sólida, com a maioria das instâncias sendo classificadas corretamente: 201 gliomas, 204 meningiomas, 207 casos sem tumor e 209 tumores pituitários, de um total de 210 amostras por classe. Contudo, a análise dos erros de classificação (elementos fora da diagonal) oferece *insights* valiosos sobre o desbalanceamento intraclasse:

- **Glioma como principal fonte de erro**: Das 210 amostras de *glioma*, 7 foram classificadas como *meningioma* e 2 como *pituitary*, totalizando 9 erros — a maior dispersão entre as classes. A confusão com meningioma decorre da semelhança textural em certas ressonâncias, onde as bordas tumorais e a densidade de massa podem mimetizar padrões de ambas as patologias no espaço latente.
- **Confiabilidade da Classe de Controle ("No Tumor")**: A classe de tecido saudável obteve um desempenho notável, com 207 acertos de 210 amostras. Apenas 3 casos foram confundidos (1 com meningioma e 2 com pituitary), e a precisão de 100% nesta classe indica que nenhuma outra patologia foi erroneamente rotulada como saudável, minimizando o risco clínico de falsos negativos.
- **Estabilidade da Classe Pituitary**: Com 209 acertos de 210 amostras (*recall* de 99,52%), a classe de adenoma pituitário demonstrou a maior facilidade

de separação, provavelmente devido à sua localização anatômica distinta e características de contraste bem definidas que o modelo conseguiu capturar com precisão.

– **Meningioma como principal receptor de confusão**: A classe *meningioma* concentrou a maior parte dos falsos positivos do modelo (7 gliomas, 1 caso sem tumor e 1 pituitary foram a ela atribuídos), o que explica sua precisão de 0,9577 — a menor entre as quatro categorias.

Essa distribuição de erros sugere que, enquanto a arquitetura é extremamente eficaz em distinguir o tecido saudável do patológico, o desafio remanescente reside no refinamento da fronteira de decisão entre gliomas e meningiomas, onde a ambiguidade visual demanda mecanismos de atenção ainda mais granulares para capturar micro-variações morfológicas.

5.2.4 Impacto da Reamostragem no Espaço Latente

Com o objetivo de avaliar isoladamente o efeito das técnicas de reamostragem sobre o domínio de imagens, aplicaram-se as estratégias SMOTE-ENN e ADASYN sobre os *embeddings* extraídos da CNN, treinando-se um classificador de referência sobre cada versão dos dados. A Tabela 6 apresenta os resultados.

Tabela 6 – Impacto das técnicas de reamostragem no espaço latente — Base MRI.

Estratégia	**Acurácia**	**F1-Score**	**G-Mean**	**AUC-ROC**
Baseline	0,8512	0,8504	0,8406	0,9737
SMOTE-ENN	0,8381	0,8347	0,8282	0,9507
ADASYN	0,8512	0,8504	0,8406	0,9737

Fonte: Elaborado pelo autor (2026).

Diferentemente do domínio financeiro, a reamostragem não produziu ganho de desempenho no domínio de imagens: o ADASYN manteve métricas idênticas à linha de base, enquanto o SMOTE-ENN provocou leve degradação (redução de 0,8406 para 0,8282 no G-Mean). Esse comportamento, à primeira vista contraintuitivo, é coerente com a natureza do *corpus* e reforça a tese central deste trabalho. O *Brain Tumor MRI Dataset* apresenta distribuição praticamente equilibrada entre as quatro classes (aproximadamente 980 amostras por categoria no conjunto de treinamento), de modo que o desbalanceamento presente neste domínio não é de natureza **interclasse**, mas **intraclasse** — manifestando-se pela variabilidade morfológica interna de cada categoria, e não pela disparidade numérica entre elas.

Como as técnicas SMOTE-ENN e ADASYN foram concebidas para corrigir o desbalanceamento interclasse, sua aplicação a um conjunto já equilibrado não encontra escassez numérica a compensar; ao contrário, a interpolação sintética

e a filtragem por vizinhança podem introduzir ruído no espaço latente, o que explica a leve degradação observada com o SMOTE-ENN. Esse resultado evidencia, de forma empírica, que o tratamento do desbalanceamento intraclasse em imagens médicas não deve ser buscado por reamostragem, mas sim pela sofisticação arquitetural — precisamente o papel desempenhado pelos mecanismos de atenção (CBAM) e pelos *transformers* hierárquicos (Swin), cujos ganhos foram demonstrados na Seção 5.2.4 anterior.

5.3 Discussão Transversal: Convergência Metodológica e o Desafio Intraclasse

A análise comparativa entre os domínios financeiro e médico revela que o desbalanceamento intraclasse, embora manifestado em naturezas de dados distintas (tabulares e imagens), impõe obstáculos análogos à convergência de modelos robustos. No domínio financeiro, a problemática centrou-se na **sobreposição estatística** em um espaço de características de alta dimensionalidade, onde o ruído presente nas transações legítimas mimetizava assinaturas fraudulentas. A solução estratégica baseou-se na técnica híbrida SMOTE-ENN, que não apenas expandiu a densidade da classe minoritária, mas refinou a topologia da fronteira de decisão via filtragem estatística. Esta abordagem, aliada ao uso do algoritmo CatBoost e sua arquitetura de árvores simétricas, permitiu uma regularização superior, mitigando o *overfitting* comum em conjuntos de dados sintetizados artificialmente.

Por outro lado, no domínio da saúde, o desafio transcendeu a disparidade numérica, manifestando-se como uma **variabilidade morfológica** complexa. Aqui, o desbalanceamento intraclasse residia na heterogeneidade anatômica do tecido cerebral saudável, que frequentemente apresentava intensidades de pixel e texturas sobrepostas a padrões tumorais sutis. A mitigação deste fenômeno não foi alcançada via reamostragem, mas sim através da sofisticação arquitetural: o uso de mecanismos de atenção (CBAM) e transformadores de visão (*Swin Transformer*). Estas tecnologias permitiram que o modelo estabelecesse dependências de longo alcance e foco seletivo em regiões críticas, "ignorando"as variações irrelevantes da classe majoritária para priorizar as características patológicas raras.

Em ambos os cenários de missão crítica, a métrica **G-Mean** consolidou-se como o indicador de desempenho mais fidedigno, superando a "falácia da acurácia nominal". Enquanto a acurácia global seria facilmente inflada pela predominância das classes majoritárias, o G-Mean — ao calcular a média geométrica entre sensibilidade e especificidade — exigiu um desempenho de excelência em ambas as frentes. Os resultados obtidos (0,9245 no domínio de fraude e 0,9773 no diagnóstico oncológico)

comprovam que a integração entre o tratamento do espaço de características e a seleção de arquiteturas com viés indutivo apropriado é a estratégia definitiva para garantir a integridade e a segurança de sistemas inteligentes diante de subgrupos complexos e dados severamente desbalanceados.

5.4 Resultados das Estratégias de Desbalanceamento

A eficácia das estratégias de reamostragem e ponderação foi o alicerce para a estabilização do aprendizado. No domínio financeiro, a aplicação do **SMOTE-ENN** não apenas equalizou o volume de amostras, mas atuou como um refinador topológico. Conforme observado na transição dos dados brutos para o conjunto balanceado, a filtragem via ENN eliminou instâncias da classe majoritária que apresentavam alta similaridade visual (no espaço de características) com as fraudes sintéticas. Isso reduziu drasticamente a ambiguidade na fronteira de decisão, permitindo que o otimizador convergisse para uma solução que prioriza a sensibilidade sem sacrificar a precisão.

Já no domínio da saúde, a estratégia divergiu para a adaptação arquitetural. Em vez de sintetizar imagens — o que poderia introduzir artefatos médicos irreais — optou-se pela reamostragem no espaço latente e pelo uso de mecanismos de atenção. A capacidade da **EfficientNet+CBAM** e do **Swin Transformer** em ignorar ruídos anatômicos da classe "No Tumor"provou ser equivalente a um processo de limpeza de dados em tempo real, atacando o desbalanceamento intraclasse através do refinamento da atenção do modelo sobre subgrupos patológicos críticos.

5.5 Comparação de Desempenho entre as Técnicas

A análise comparativa revela uma distinção clara entre a eficácia dos modelos conforme a natureza dos dados. Para os dados tabulares transformados por PCA, o paradigma de *Gradient Boosting*, liderado pelo **CatBoost**, demonstrou superioridade sobre redes neurais densas e outros algoritmos de *ensemble*. A robustez das árvores simétricas do CatBoost em lidar com o desbalanceamento intraclasse manifestou-se na menor variância entre as dobras da validação cruzada, mantendo um G-Mean de 0,9245.

Tabela 7 – Comparação de G-Mean entre estratégias de balanceamento (Média $k = 5$)

Estratégia	**XGBoost**	**LightGBM**	**CatBoost**	**Árvore**
SMOTE-ENN	0,9214	**0,9254**	0,9245	—
HEM	0,8822	0,7262	**0,8960**	0,8786

Fonte: Elaborado pelo autor (2026).

Em contrapartida, no processamento de imagens MRI, a hierarquia dos *Transformers* superou as Convoluções tradicionais. O **Swin Transformer** (Acurácia: 97,74%) demonstrou uma capacidade superior de capturar dependências globais, algo essencial quando o tumor apresenta características morfológicas dispersas. Enquanto o CatBoost resolveu a sobreposição estatística via regularização rigorosa, o Swin Transformer resolveu a variabilidade visual via auto-atenção local e janelas deslocadas, provando que arquiteturas especializadas são mandatórias em cenários de alta complexidade.

5.6 Análise das Métricas e Impacto do Desbalanceamento Intraclasse

O foco central desta pesquisa foi o impacto do desbalanceamento intraclasse — fenômeno onde subgrupos da classe majoritária mimetizam a classe minoritária. A utilização da acurácia nominal mostrou-se insuficiente, pois ocultaria falhas críticas. A métrica **G-Mean** consolidou-se como o "pilar de justiça"do modelo: atingir 0,9245 na fraude e 0,9773 no tumor indica que os modelos não apenas aprenderam a classe dominante, mas capturaram a essência estatística das minorias.

O desbalanceamento intraclasse foi o principal motor de erro identificado. Na detecção de fraude, o modelo confundiu transações legítimas de alto valor com fraudes (sobreposição de magnitude). No diagnóstico médico, a confusão ocorreu entre gliomas e meningiomas devido a texturas similares. A análise prova que o problema não é apenas a "quantidade"de dados (desbalanceamento interclasse), mas a "qualidade"da separação entre subgrupos complexos (desbalanceamento intraclasse), o que justifica o uso de técnicas híbridas e modelos baseados em atenção.

5.7 Limitações e Possíveis Fontes de Erro

Apesar dos resultados de alta performance, esta pesquisa identifica limitações inerentes. No domínio financeiro, o uso de dados sintetizados via SMOTE pode induzir um viés de "linearidade excessiva"na fronteira de decisão, uma vez que o algoritmo interpola instâncias existentes. Além disso, a opacidade dos atributos PCA impede uma interpretação semântica direta sobre quais variáveis do mundo real (ex: localização ou horário) são mais suscetíveis à fraude.

No domínio médico, a principal limitação reside no viés de aquisição dos dados de MRI. Artefatos de movimento ou variações no contraste do equipamento podem ser interpretados pela rede como características patológicas, o que explica os erros residuais de classificação entre meningiomas e gliomas observados na Matriz de

Confusão. Por fim, o custo computacional para o treinamento de modelos como o Swin Transformer é significativamente superior às CNNs tradicionais, o que deve ser considerado em implementações de sistemas de tempo real em dispositivos médicos de ponta.

6
Conclusão

Este capítulo encerra a presente pesquisa, sintetizando as deduções obtidas ao longo do desenvolvimento experimental e avaliando o cumprimento dos objetivos propostos. Diante dos desafios inerentes ao aprendizado de máquina em cenários de alta complexidade e assimetria de dados, as seções a seguir consolidam as considerações finais, as contribuições técnicas e científicas geradas, bem como as diretrizes para investigações futuras.

6.1 Considerações Finais

O desbalanceamento de classes, sobretudo em sua vertente intraclasse, constitui uma das barreiras mais severas para a confiabilidade de sistemas inteligentes em ambientes de missão crítica. Esta pesquisa propôs e validou um arcabouço metodológico bimodal, investigando frentes distintas da ciência de dados: o domínio financeiro tabular e o domínio médico baseado em visão computacional. A convergência dos resultados obtidos demonstra que a mitigação desse fenômeno não reside em soluções genéricas, mas sim na sinergia entre o tratamento rigoroso do espaço de características e a seleção de arquiteturas dotadas de viés indutivo apropriado.

No cenário financeiro, o diagnóstico inicial revelou um abismo estatístico crônico (0,172% de fraudes), onde sinais preditivos fortes eram sistematicamente sufocados pela densidade da classe majoritária. A implementação da estratégia híbrida SMOTE-ENN provou ser definitiva ao redefinir a topologia das fronteiras de decisão por meio da síntese adaptativa e filtragem estatística de ruídos. Esse tratamento viabilizou o desempenho robusto dos algoritmos de *Gradient Boosting*, com destaque para a consistência do CatBoost (G-Mean de 0,9245), tecnicamente empatado com o XGBoost (0,9214) e o LightGBM (0,9254). Esse tratamento viabilizou o desempenho robusto dos algoritmos de *Gradient Boosting*, com destaque para a consistência do CatBoost (G-Mean de 0,9245), tecnicamente empatado com o XGBoost (0,9214) e o LightGBM (0,9254). A arquitetura de árvores simétricas do classificador agiu como um regularizador natural, neutralizando a variância introduzida por dados sintéticos e mitigando com sucesso a sobreposição estatística de transações legítimas complexas.

Paralelamente, no cenário médico de classificação de tumores cerebrais, constatou-se que o desbalanceamento intraclasse manifesta-se pela heterogeneidade morfológica e

textural de tecidos saudáveis na classe majoritária. Em vez do caminho convencional de reamostragem sintética, que poderia comprometer a semântica biológica das imagens, a solução fundamentou-se na sofisticação arquitetural. O modelo Swin Transformer, baseado na hierarquia de janelas deslocadas e auto-atenção local, alcançou o estado da arte com uma acurácia global de 97,74% e um Macro F1-Score de 0,9775. A capacidade da rede em estabelecer dependências espaciais de longo alcance permitiu discernir com precisão cirúrgica assinaturas patológicas de baixo contraste, minimizando os falsos positivos na classe saudável de controle. Em suma, a substituição da acurácia nominal pela Média Geométrica (G-Mean) como métrica central de auditoria confirmou que ambos os sistemas atingiram um equilíbrio operacional seguro, preservando a integridade e a sensibilidade de detecção das instâncias raras.

6.2 Contribuições do Trabalho

A principal contribuição técnico-científica originada por este trabalho compreende o desenvolvimento de um framework metodológico unificado e estruturado. Este fluxo experimental documentado serve como diretriz para o enfrentamento do desbalanceamento intraclasse, demonstrando aplicabilidade tanto para dados tabulares contínuos submetidos a transformações lineares (PCA) quanto para tensores de alta resolução representados por imagens MRI. O estudo preenche uma lacuna relevante na literatura de engenharia ao prover uma análise comparativa rigorosa de variância e desvio padrão entre algoritmos de *Gradient Boosting* sob o efeito de reamostragem sintética híbrida, consolidando a superioridade do CatBoost em termos de estabilidade computacional e generalização estatística.

Adicionalmente, o trabalho contribui significativamente para o campo da visão computacional médica ao demonstrar o impacto prático da integração de módulos de atenção convolucional e transformadores hierárquicos. Os resultados alcançados pelo Swin Transformer validam empiricamente que a sofisticação da arquitetura na captura de dependências contextuais globais supera as limitações de kernels convolucionais locais, estabelecendo novos patamares de acurácia no diagnóstico de patologias sobrepostas sem a necessidade de expansão artificial do banco de imagens. Por fim, a pesquisa promove uma ruptura com a falácia da acurácia nominal em sistemas de missão crítica, consolidando o uso de métricas multivariadas como o G-Mean e o Macro F1-Score para a auditoria de sistemas de inteligência artificial de alta responsabilidade.

6.3 Sugestões para Trabalhos Futuros

Apesar dos resultados alcançados, este trabalho abre portas para diversas investigações futuras. Um primeiro passo natural seria incorporar técnicas de Inteligência Artificial Explicável (XAI). No cenário financeiro, a aplicação do método SHAP ajudaria a abrir a "caixa preta"criada pelo PCA, permitindo entender de forma clara o peso de cada variável original na decisão final do modelo. Já na área médica, o uso de algoritmos como o Grad-CAM, devidamente adaptados para arquiteturas de atenção, geraria mapas de ativação visual muito mais precisos. Isso daria aos especialistas a chance de validar clinicamente se o Swin Transformer está, de fato, focando nas regiões patológicas corretas durante as predições.

Outro ponto que merece ser aprofundado antes mesmo da modelagem é o diagnóstico prévio dos dados. Sugere-se o uso de métricas matemáticas de complexidade, como as clássicas medidas de Ho e Basu (a exemplo da fração de pontos na fronteira de decisão, N_1, e do volume da região de sobreposição, F_2). Medir esses fatores permitiria entender o grau real de *overlap* e a presença de pequenos grupos isolados (*small disjuncts*) nas classes minoritárias. Conhecendo a fundo essa geometria da base, seria o momento ideal para evoluir técnicas como o SMOTE-ENN para métodos guiados por agrupamento, como o *K-Means SMOTE* ou o *PC-SMOTE*. Essas abordagens evitam gerar dados artificiais em áreas muito misturadas, lidando muito melhor com a fragmentação intraclasse e com agrupamentos de densidade variável.

Olhando novamente para as imagens médicas, um fenômeno prático e crítico que precisa ser investigado é a "Estratificação Oculta"(*Hidden Stratification*). Na prática, isso ocorre quando o modelo falha bastante em subgrupos muito específicos de pacientes (como tumores visualmente sutis), mas que não estão rotulados separadamente na base original. Para mitigar esse viés sem a necessidade de anotar manualmente milhares de imagens, uma ótima saída seria testar algoritmos de Otimização Robusta, como o *Group DRO* ou o *Just Train Twice* (JTT). O JTT, por exemplo, foca em treinar e melhorar o desempenho justamente nas amostras em que o modelo padrão mais erra, otimizando de forma indireta o chamado pior grupo. Se combinarmos isso com funções de perda focadas na dificuldade da amostra, a própria rede passaria a tratar o desbalanceamento intraclasse de forma nativa.

Para a parte de geração artificial de imagens, explorar Redes Adversariais Generativas (GANs) é um caminho clássico, mas a literatura mais recente aponta que os Modelos de Difusão (*Diffusion Models*) conseguem ir além. Técnicas baseadas em difusão contornam a instabilidade e o colapso de modo (*mode collapse*) que costumam atrapalhar as GANs na hora de sintetizar patologias muito raras (como o *DiffMix*, desenhado especificamente para imagens médicas). A ideia seria testar se gerar

imagens com altíssima fidelidade consegue replicar, no espaço visual, o mesmo nível de sucesso que tivemos com o SMOTE-ENN nos dados tabulares.

Por fim, saindo um pouco do foco puramente preditivo e indo para a engenharia de software, sabemos que modelos como os transformadores de visão têm um custo computacional altíssimo. Seria muito interessante aplicar técnicas de quantização de pesos e destilação de conhecimento para deixá-los mais leves. Isso viabilizaria rodar essas inteligências diretamente em dispositivos de borda (*edge computing*), aproximando a nossa solução de aplicações em tempo real, seja num equipamento clínico local ou num sistema financeiro com restrição severa de recursos. Em paralelo a isso, tentar modelar as transações bancárias usando Redes Neurais de Grafos (GCN) surge como uma excelente aposta final para capturar as conexões entre as transações de forma bem mais resistente ao desbalanceamento de dados.

Referências

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. In: *SIGKDD Explorations*. [S.l.: s.n.], 2004. v. 6, n. 1, p. 20–29. Citado na página 30.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado 2 vezes nas páginas 18 e 19.

BRAIN Tumor MRI Dataset. 2025. Conjunto de dados contendo 7.023 imagens de ressonância magnética. Citado 3 vezes nas páginas 34, 35 e 36.

BRANCO, P.; TORGO, L.; RIBEIRO, R. P. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, v. 49, n. 2, p. 31–40, 2016. Citado 5 vezes nas páginas 24, 27, 34, 39 e 45.

BUDA, M.; MAKI, A.; MAZUROWSKI, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, v. 106, p. 249–259, 2018. Citado na página 28.

BUREZ, J.; POEL, D. V. d. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, v. 36, n. 3, p. 4626–4636, 2009. Citado na página 35.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Citado na página 30.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002. Citado na página 38.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, p. 785–794, 2016. Citado na página 23.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, p. 785–794, 2016. Citado na página 40.

DOMINGOS, P. *A few useful things to know about machine learning*. [S.l.]: ACM, 2012. v. 55. 78–87 p. Citado na página 18.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: unbiased boosting with categorical features. *arXiv preprint arXiv:1810.11363*, 2018. Citado 2 vezes nas páginas 23 e 24.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems (NeurIPS)*, p. 6638–6648, 2018. Citado 2 vezes nas páginas 40 e 41.

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 26.

FERNáNDEZ, A.; GARCIA, S.; HERRERA, F.; CHAWLA, N. V. Learning from imbalanced data sets. *ACM Computing Surveys*, v. 52, n. 4, p. 1–36, 2018. Citado 5 vezes nas páginas 19, 24, 27, 29 e 30.

FERNáNDEZ, A.; GARCíA, S.; HERRERA, F.; CHAWLA, N. V. Learning from imbalanced data sets. *ACM Computing Surveys*, v. 52, n. 4, p. 1–36, 2018. Citado 3 vezes nas páginas 35, 38 e 45.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: O'Reilly Media, 2019. Citado na página 18.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Citado na página 18.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. [S.l.]: Springer, 2009. Citado na página 18.

HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: IEEE. *2008 IEEE International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2008. p. 1322–1328. Citado 3 vezes nas páginas 31, 38 e 39.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263–1284, 2009. Citado na página 23.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, v. 21, n. 9, p. 1263–1284, 2009. Citado 3 vezes nas páginas 34, 38 e 45.

JOLLIFFE, I. T. *Principal Component Analysis*. [S.l.]: Springer, 2002. Citado na página 35.

KE, G.; MENG, Q.; FINLEY, T.; AL. et. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NeurIPS)*, p. 3146–3154, 2017. Citado 2 vezes nas páginas 23 e 24.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, p. 3146–3154, 2017. Citado na página 40.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 2015. Citado na página 20.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, v. 18, n. 1, p. 559–563, 2017. Citado 2 vezes nas páginas 35 e 38.

LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; GUO, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2021. p. 10012–10022. Citado 3 vezes nas páginas 22, 36 e 44.

LIU, Z.; WEI, Z.; TIAN, Y. Handling intra-class imbalance in deep learning models. *IEEE Transactions on Neural Networks and Learning Systems*, v. 33, n. 7, p. 3129–3141, 2022. Citado na página 28.

MITCHELL, T. *Machine Learning*. [S.l.]: McGraw-Hill, 1997. Citado na página 18.

MOUSAVI, S.; GHASEMIAN, A.; SHAMSOLMOALI, P. Class imbalance problem in medical diagnosis: A review. *Medical Data Analysis*, v. 27, n. 3, p. 112–134, 2020. Citado na página 28.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. [S.l.]: MIT Press, 2012. Citado na página 18.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, v. 2, n. 1, p. 37–63, 2011. Citado na página 24.

PROVOST, F.; FAWCETT, T.; KOHAVI, R. The case against accuracy estimation for comparing induction algorithms. *ICML*, v. 98, p. 445–453, 1998. Citado na página 25.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986. Citado na página 19.

SHRIVASTAVA, A.; GUPTA, A.; GIRSHICK, R. Training region-based object detectors with online hard example mining. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 761–769. Citado na página 31.

TAN, M.; LE, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. *International conference on machine learning*. [S.l.], 2019. p. 6105–6114. Citado 2 vezes nas páginas 21 e 36.

ULB. *Credit Card Fraud Detection Dataset*. 2016. Disponível em: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>. Citado 2 vezes nas páginas 34 e 35.

WILSON, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, IEEE, n. 3, p. 408–421, 1972. Citado 3 vezes nas páginas 30, 31 e 38.

WOO, S.; PARK, J.; LEE, J.-Y.; KWEON, I. S. Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2018. p. 3–19. Citado 2 vezes nas páginas 21 e 36.

ZHU, X.; GOLDBERG, A. B. *Introduction to Semi-Supervised Learning*. [S.l.]: Morgan & Claypool, 2009. v. 3. 1–130 p. (Synthesis Lectures on Artificial Intelligence and Machine Learning, 1). Citado na página 19.