# Hy-Synergy: geração sintética de dados tabulares em saúde guiada por diagnóstico local

**Mauro Henrique Lima de Boni**[1,2], **Iwens Gervásio Sene Junior**[2], **Ronaldo Martins da Costa** [2]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Tocantins – IFTO

[2]Universidade Federal de Goiás – UFG
Instituto de Informática, Goiânia, Goiás. Brasil

mauro@ifto.edu.br,iwens@ufg.br,ronaldocosta@ufg.br

***Abstract.*** *Imbalanced and sensitive clinical tabular datasets hinder the training of robust predictive models. This paper presents Hy-Synergy, a modular augmentation framework based on topological diagnosis, clusters discovery, local decision policy, and routed synthetic generation. Evaluated on the Pima Indians Diabetes and Adult Census Income datasets, Hy-Synergy achieved the best trade-off among fidelity, utility, and privacy in both domains, reaching C2ST of 0.5478, JSD of 0.0312, ROC AUC of 0.8270, and MIA of 0.5242 on Pima. Results suggest that locally guided interventions are promising for expanding small and sensitive tabular datasets in health-related settings.*

***Resumo.*** *Dados tabulares clínicos desbalanceados e sensíveis dificultam o treinamento de modelos preditivos robustos. Este trabalho apresenta o Hy-Synergy, um framework modular de aumento de dados baseado em diagnóstico topológico, descoberta de clusters, política local de decisão e geração sintética roteada. Avaliado nos datasets Pima Indians Diabetes e Adult Census Income, o método apresentou o melhor equilíbrio entre fidelidade, utilidade e privacidade nos dois domínios, alcançando C2ST de 0.5478, JSD de 0.0312, ROC AUC de 0.8270 e MIA de 0.5242 no Pima. Os resultados sugerem que intervenções orientadas por estrutura local são promissoras para expansão de bases clínicas pequenas e sensíveis.*

## 1. Introdução

Dados sintéticos tabulares são frequentemente utilizados para mitigar escassez amostral, desbalanceamento e restrições de privacidade em aprendizado de máquina. Em dados clínicos, esse problema é particularmente sensível, pois a síntese precisa preservar padrões estatísticos relevantes para predição sem reproduzir ruído ou memorizar registros reais. A revisão sistemática de (de Boni et al. 2025) mostra que utilidade, fidelidade e privacidade formam os principais eixos de avaliação, embora ainda persista forte heterogeneidade metodológica e pouca padronização experimental.

Este trabalho apresenta o *Hy-Synergy*, um framework de geração sintética tabular guiado por diagnóstico espacial. Em vez de tratar a distribuição da minoria como homogênea, o método identifica disjuntos locais, também chamados de *clusters* ou subconceitos, estima dificuldade local, aplica limpeza seletiva, realiza *undersampling* direcionado da maioria e roteia a geração sintética de forma localizada. O objetivo é melhorar o

compromisso entre fidelidade, utilidade e privacidade em bases desbalanceadas, com foco principal em dados de saúde e avaliação complementar em um domínio tabular misto.

## 2. Trabalhos relacionados

A literatura de geração sintética tabular abrange reamostragem clássica e modelos profundos. Segundo a revisão de (de Boni et al. 2025), embora escassez, desbalanceamento e privacidade motivem essas técnicas, prevalecem a heterogeneidade experimental e a negligência à privacidade em prol da utilidade. Métodos como *SMOTE* e *ADASYN* focam na vizinhança geométrica para classificação desbalanceada, enquanto sintetizadores como *CTGAN* e *TVAE* buscam aprender a distribuição global de forma unificada (Panfilo et al. 2023, Alshantti et al. 2024, Apellaniz et al. 2024).

O *Hy-Synergy* distingue-se ao propor uma abordagem regionalizada, contrastando com a reamostragem de vizinhança fixa e a convergência global de modelos profundos, que frequentemente negligenciam micro-conceitos da minoria. Utilizando *Bayesian Gaussian Mixture* (BGM), o *framework* decompõe a topologia em subconceitos latentes para aplicar um roteamento inteligente. Diferente de políticas uniformes, cada *cluster* recebe intervenções específicas, como limpeza seletiva ou *undersampling* direcionado, baseadas no diagnóstico de tipologia de exemplos (seguros, *borderline* ou ruidosos), baseado em (Smith et al. 2014). Essa orquestração adapta a síntese à dificuldade estrutural de cada região, otimizando o balanço entre utilidade e privacidade.

No domínio da saúde, tais desafios são críticos, pois a qualidade estrutural dos dados impacta diretamente a utilidade preditiva e o risco de reidentificação (Inan et al. 2023, Wang and Pai 2023, Rodriguez-Almeida et al. 2023, Eckardt et al. 2023, Kang et al. 2023). O *Hy-Synergy* endereça essa lacuna ao integrar a descoberta de *clusters* ao diagnóstico local da minoria (Napierala and Stefanowski 2016). A arquitetura permite uma geração adaptada à complexidade do espaço latente, validada sob uma avaliação multidimensional que contempla, de forma integrada, métricas de utilidade, fidelidade e privacidade.

## 3. Método

O Hy-Synergy é um framework modular de geração sintética tabular guiado por diagnóstico espacial. Sua cadeia de processamento segue a sequência: *diagnose* $\rightarrow$ *policy* $\rightarrow$ *cleaning* $\rightarrow$ *undersample* $\rightarrow$ *generate* $\rightarrow$ *post-process* $\rightarrow$ *evaluate*. A proposta parte do princípio de que a geração sintética em bases desbalanceadas deve ser condicionada à estrutura local da classe minoritária e à geometria da fronteira entre classes.

### 3.1. Pré-processamento e prevenção de vazamento

O fluxo de processamento foi estruturado para evitar vazamento de informação (*data leakage*) entre treinamento e teste. Para isso, utilizou-se um método próprio, responsável pela inferência automática dos papéis das colunas, incluindo variáveis contínuas, ordinais e categóricas. Em cada dobra da validação cruzada, o ajuste (*fit*) dos transformadores foi realizado exclusivamente no conjunto de treinamento. Esse processo incluiu normalização via *StandardScaler*, transformações logarítmicas para redução de assimetria e *target encoding* para atributos categóricos. O conjunto de teste permaneceu isolado durante toda a etapa de ajuste, sendo apenas transformado com base nos parâmetros aprendidos no treinamento. Adicionalmente, foi adotado um limiar de raridade de $0{,}01$ para o tratamento

de categorias pouco frequentes, bem como a exclusão de colunas redundantes conforme os esquemas de dados definidos.

### 3.2. Diagnóstico topológico multivariado

A segunda etapa realiza o diagnóstico da classe minoritária. Aplica-se inicialmente PCA para reduzir ruído e instabilidade na estimação de covariâncias em alta dimensão. Em seguida, um modelo *Bayesian Gaussian Mixture* (BGM) identifica os *clusters* internos à minoria em um espaço latente mais compacto. Por fim, um algoritmo baseado em *k-nearest neighbors* estima a dificuldade local e rotula instâncias como *safe*, *borderline*, *rare* ou *outlier*, baseado na caracterização apresentada por (Napierala and Stefanowski 2016), informação posteriormente utilizada pela política de decisão.

### 3.3. Política de decisão e reamostragem cirúrgica

Com base nas estatísticas de cada *cluster*, o framework aplica uma política de decisão $\pi(s)$ para determinar o tipo de intervenção e sua intensidade. A política opera sobre um vetor de estado local composto por três quantidades: raridade do *cluster* ($r_s$), dureza média ($h_s$) e pressão de fronteira ($\beta_s$). Intuitivamente, $r_s$ mede a escassez relativa do *cluster*, $h_s$ resume a dificuldade local de classificação e $\beta_s$ quantifica a presença de instâncias majoritárias na vizinhança imediata da minoria.

A decisão é formulada por escores para três estratégias: sobreamostragem, subamostragem e intervenção híbrida. O escore de sobreamostragem prioriza *clusters* raros e difíceis,

$$score^{s}_{over} = \alpha_1 r_s + \alpha_2 \min\left(h_s, q_{0.90}(h)\right),$$

enquanto o escore de subamostragem favorece regiões com maior pressão da classe majoritária,

$$score^{s}_{under} = \alpha_3(1 - r_s) + \alpha_4 \beta_s.$$

Para zonas de conflito, define-se ainda um escore híbrido,

$$score^{s}_{hyb} = \lambda \cdot \min\left(score^{s}_{over}, score^{s}_{under}\right),$$

o qual representa um compromisso controlado entre síntese e remoção seletiva.

A ação final é escolhida por

$$a_s = \arg\max\{score^{s}_{over},\ score^{s}_{under},\ score^{s}_{hyb},\ 0\},$$

de modo que a política também pode optar por não intervir quando nenhum escore ultrapassa o estado nulo. A intensidade da intervenção é então normalizada por escalonamento por quantis,

$$\tilde{w}_s = \text{quantile\_scale}(score^{s}_{a_s}; q_{0.25}, q_{0.75}, w_{min}, w_{max}),$$

mapeando os escores brutos para uma faixa operacional de reamostragem. Na prática, essa formulação permite que o *Hy-Synergy* combine, de forma interpretável, limpeza seletiva da minoria, *undersampling* cirúrgico da maioria e síntese localizada em *clusters* nos quais a intervenção é mais justificável.

**Tabela 1. Hiperparâmetros da política $\pi(s)$ (valores padrão e descrição).**

| Parâmetro | Valor padrão | Descrição |
|---|---|---|
| $\alpha_1, \alpha_2$ | 0.5, 0.5 | Pesos de raridade ($r_s$) e dureza ($h_s$) no $score_{over}$ |
| $\alpha_3, \alpha_4$ | 0.5, 0.5 | Pesos de não raridade e pressão de fronteira ($\beta_s$) no $score_{under}$ |
| $\lambda$ | 0.75 | Fator de fusão da estratégia híbrida |
| $w_{min}, w_{max}$ | 0.05, 0.40 | Limites da intensidade de reamostragem por cluster |
| $\tau_u$ | 0.10 | Gatilho mínimo para efetivar *undersampling* |
| $\theta$ | 0.25 | Fração mínima de instâncias ruidosas para acionar *oversampling* |

### 3.4. Geração roteada e pós-processamento

Após a preparação do espaço de aprendizagem, a geração sintética é roteada para os *clusters* ativos. Essa decisão evita que a síntese seja realizada de forma indiferenciada sobre a distribuição global. Em seguida, um pós-processador estatístico ajusta covariâncias, momentos univariados e, quando necessário, a aderência à grade dos dados reais. O objetivo é reduzir discrepâncias marginais e conjuntas sem introduzir artefatos adicionais.

## 4. Desenho experimental

A avaliação foi conduzida com os datasets *Pima Indians Diabetes* (Smith et al. 1988) e *Adult Census Income* (Becker and Kohavi 1996), selecionados por representarem cenários distintos de geração tabular. O primeiro oferece um domínio predominantemente contínuo, com aplicação biomédica. O segundo representa um domínio misto, com variáveis numéricas e categóricas de alta cardinalidade, além de desbalanceamento mais acentuado. No conjunto *Adult*, a classe positiva corresponde à faixa de renda superior a 50K, tratada como minoria.

### 4.1. Configuração experimental

Para garantir estabilidade estatística e reprodutibilidade, os experimentos foram conduzidos por meio de validação cruzada repetida. Foram utilizadas 5 sementes (*seeds*) distintas para o embaralhamento dos dados e, para cada semente, aplicou-se um *Stratified K-Fold* com 5 dobras (*splits*). Esse arranjo resultou em 25 execuções independentes para cada configuração de gerador e dataset. A estratificação foi empregada para preservar a proporção da classe minoritária em todas as partições de treinamento e teste, reduzindo distorções em métricas sensíveis ao desbalanceamento, como F1-score e ROC-AUC.

### 4.2. Protocolos de avaliação

Foram adotados três protocolos de avaliação. No protocolo **TRTR**, os modelos são treinados e testados em dados reais, funcionando como linha de base empírica. No protocolo **HybridTR**, os modelos são treinados em uma combinação de dados reais e sintéticos e avaliados em teste real. No protocolo **TSTR**, o treinamento utiliza apenas dados sintéticos, enquanto a avaliação ocorre em dados reais não vistos.

### 4.3. Baselines e modelos downstream

Os métodos de comparação foram escolhidos para cobrir famílias representativas identificadas na revisão sistemática de literatura (de Boni et al. 2025), que evidencia recorrência de técnicas clássicas de reamostragem, variantes baseadas em GANs, modelos VAE adaptados para dados tabulares e métodos estatísticos baseados em copulas. Por essa razão, o estudo inclui métodos geométricos clássicos, como *SMOTE*

(Chawla et al. 2002), *ADASYN* (He et al. 2008), *Borderline-SMOTE* (Han et al. 2005) e *SVMSMOTE* (Nguyen et al. 2011), além de sintetizadores profundos amplamente recorrentes na literatura recente, como *CTGAN* e *TVAE* (Xu et al. 2019). Em análises complementares, também foi considerado o *Gaussian Copula* (Patki et al. 2016) como baseline estatístico.

Os sintetizadores profundos *CTGAN* e *TVAE* foram configurados sob a perspectiva de modelagem global da distribuição, para contraste com a abordagem regionalizada do *Hy-Synergy*. A utilidade preditiva final foi avaliada por quatro algoritmos *downstream* — *LightGBM*, *XGBoost*, *Random Forest* e Regressão Logística — a fim de mitigar o viés de dependência arquitetural de um único classificador.

### 4.4. Otimização de hiperparâmetros

A otimização de hiperparâmetros do *framework Hy-Synergy* foi conduzida de forma autônoma via Optuna, empregando o algoritmo evolucionário NSGA-II com 40 iterações (*trials*) por dobra da validação cruzada. Diferentemente de otimizações tradicionais focadas apenas em performance preditiva, a busca foi modelada como um problema multiobjetivo visando a descoberta de uma Fronteira de Pareto ótima entre Utilidade (e.g., ROC AUC e F1-Score), Fidelidade topológica e Privacidade dos dados gerados. O espaço de busca para a modelagem do espaço latente contemplou: (i) a taxa de variância retida na etapa de compressão via PCA (*pca_components* $\in [0.85, 0.99]$); (ii) o limite superior de partições BGM ($max\_clusters \in [4, 25]$); e (iii) a penalização bayesiana contra fragmentação excessiva, controlada pelo *Dirichlet Prior* ($weight\_concentration\_prior \in [0.01, 0.99]$ em escala logarítmica). Para a política dinâmica de reamostragem ($\pi$), os pesos heurísticos de intervenção topológica ($\alpha_1, \alpha_2, \alpha_3, \alpha_4$) foram avaliados no intervalo de $[0.1, 1.0]$. Por fim, ressalta-se que, em contraste com os *baselines* geométricos que dependem da busca empírica pelo hiperparâmetro de vizinhança ($k$), o diagnóstico estrutural do *Hy-Synergy* é fundamentalmente probabilístico: ele dispensa definições espaciais rígidas, adaptando-se organicamente à estrutura dos dados de forma autônoma através das matrizes de covariância do BGM.

Visando uma comparação equânime, os hiperparâmetros das *baselines* foram otimizados via busca Bayesiana, preterindo-se o uso de configurações padrão da literatura. Em métodos geométricos, ajustou-se o número de vizinhos (k) à densidade local de cada base; para modelos profundos (*CTGAN* e *TVAE*), otimizaram-se parâmetros de arquitetura e treinamento, como *learning rate*, *batch size* e épocas. Esse rigor metodológico mitiga riscos de subajuste ou colapso de modo inerentes a configurações genéricas, estabelecendo um referencial competitivo para a validação do Hy-Synergy.

Para a etapa de rotulagem de instâncias, descrita na seção 3.2, fixou-se k=5 para o diagnóstico de vizinhança. Tal valor atua apenas como uma janela de observação local estável para a classificação de tipos de exemplos, não interferindo na partição topológica principal, que permanece sob a responsabilidade das distribuições probabilísticas do BGM.

### 4.5. Métricas e análise estatística

A utilidade foi avaliada via *ROC-AUC* e *F1-Optimal*, a fidelidade por *C2ST* e *JSD*, e a privacidade por *MIA Accuracy*. Métricas complementares (*Log Loss*, *TSTR*, *DCR* e *MMD*) foram analisadas, mas omitidas das tabelas principais por restrição de espaço.

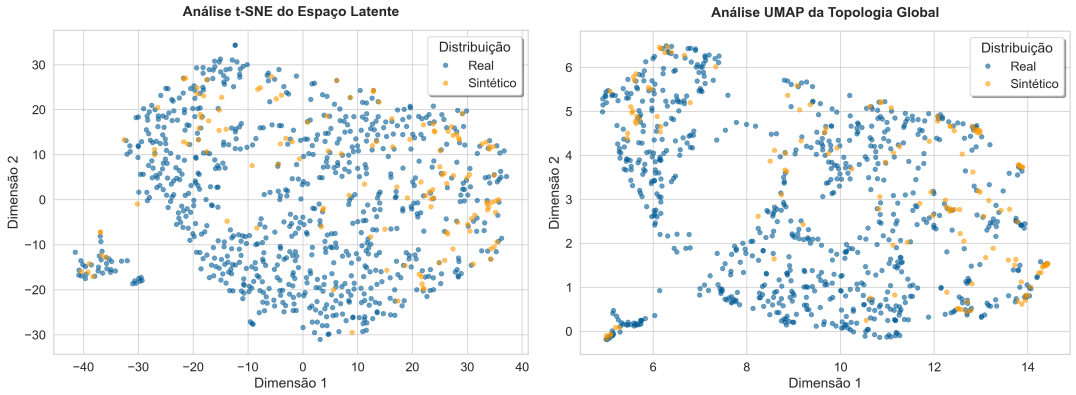


**Figura 1. Projeções bidimensionais por t-SNE e UMAP comparando dados reais (azul) e sintéticos gerados pelo *Hy-Synergy* (laranja) - dataset *Pima*.**

Seguindo o protocolo de Demšar (Demšar 2006), aplicou-se o teste de *Friedman* para verificar diferenças globais, seguido pelo teste *post-hoc* de *Nemenyi* com diagramas de diferença crítica ($p < 0{,}05$) para identificação de pares significativos. Para o estudo de ablação (versões sem política $\pi(s)$, *undersampling*, *cleaning* ou *post-processing*), empregou-se o teste de postos sinalizados de *Wilcoxon*, preservando o alinhamento de amostras por semente e dobra. Todas as comparações múltiplas incluíram controle de erro do tipo I.

## 5. Resultados

Os resultados são apresentados por dataset e por protocolo de avaliação. No dataset *Pima*, a análise enfatiza a preservação de estrutura contínua e a utilidade preditiva em um contexto biomédico. No dataset *Adult*, avalia-se a robustez da proposta em um domínio misto, com variáveis categóricas de maior cardinalidade e maior complexidade combinatória. A análise qualitativa é utilizada apenas como evidência exploratória complementar, enquanto as conclusões principais se apoiam nas métricas quantitativas de fidelidade, utilidade e privacidade. Por restrição de espaço, as Tabelas 2–7 apresentam os métodos principais reportados no desenho experimental, enquanto resultados complementares com Gaussian Copula permaneceram restritos às análises auxiliares.

### 5.1. Avaliação qualitativa do *Hy-Synergy*

Antes da comparação quantitativa, realizou-se uma inspeção visual dos dados sintéticos por meio de t-SNE e UMAP. As Figuras 1 e 2 comparam amostras reais e sintéticas nos dois domínios analisados.

No dataset *Pima*, observa-se sobreposição visual consistente entre amostras reais e sintéticas em diferentes agrupamentos locais. No dataset *Adult*, a cobertura é menos homogênea, mas as amostras sintéticas ainda se concentram em regiões compatíveis com a estrutura dos dados reais. Essas projeções devem ser interpretadas apenas como evidência exploratória e são complementadas, nas subseções seguintes, pelas métricas quantitativas de fidelidade, utilidade e privacidade.

### 5.2. Fidelidade

As Tabelas 2 e 3 apresentam os resultados de fidelidade para os datasets *Pima* e *Adult*. Nesta análise, o C2ST quantifica a separabilidade entre amostras reais e sintéticas, de

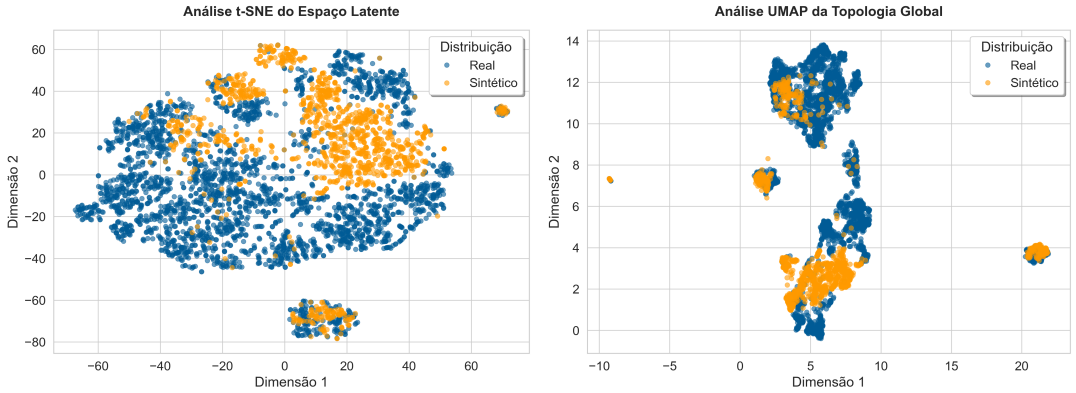


**Figura 2. Projeções bidimensionais por t-SNE e UMAP comparando dados reais (azul) e sintéticos gerados pelo *Hy-Synergy* (laranja) - dataset *Adult*.**

modo que valores mais próximos de 0.5 indicam maior indistinguibilidade entre os conjuntos. A JSD mede a divergência entre distribuições marginais, sendo desejáveis valores menores. Em conjunto, essas métricas permitem avaliar em que medida os dados sintéticos preservam simultaneamente a estrutura global e o comportamento univariado das variáveis, em consonância com os eixos de avaliação discutidos pela revisão sistemática de (de Boni et al. 2025).

**Tabela 2. Fidelidade sintética no dataset Pima. Para o C2ST, o valor ideal é 0.5; para a JSD, valores menores indicam melhor alinhamento marginal**

| Método | C2ST | JSD (↓) |
|---|---|---|
| **Hy-Synergy** | $0.5478 \pm 0.0456$ | $0.0312 \pm 0.0033$ |
| ADASYN | $0.8640 \pm 0.0242$ | $0.0433 \pm 0.0037$ |
| SVMSMOTE | $0.8642 \pm 0.0159$ | $0.0462 \pm 0.0043$ |
| SMOTE | $0.8723 \pm 0.0213$ | $0.0514 \pm 0.0045$ |
| Borderline-SMOTE | $0.9018 \pm 0.0154$ | $0.0475 \pm 0.0048$ |
| TVAE | $0.9806 \pm 0.0076$ | $0.1429 \pm 0.0201$ |
| CTGAN | $0.9857 \pm 0.0076$ | $0.1418 \pm 0.0242$ |

No dataset Pima, o Hy-Synergy apresentou o melhor desempenho nas duas métricas, com **C2ST =** $0.5478 \pm 0.0456$, próximo do ideal teórico e **JSD =** $0.0312 \pm 0.0033$. Em contraste, CTGAN e TVAE produziram C2ST superior a 0.98, indicando alta separabilidade entre dados reais e sintéticos, enquanto os métodos geométricos exibiram comportamento intermediário.

**Tabela 3. Fidelidade sintética no dataset Adult. Para o C2ST, o valor ideal é 0.5; para a JSD, valores menores indicam melhor alinhamento marginal**

| Método | C2ST | JSD (↓) |
|---|---|---|
| Hy-Synergy | $0.8844 \pm 0.0089$ | $0.0125 \pm 0.0020$ |
| SVMSMOTE | $0.9621 \pm 0.0242$ | $0.0675 \pm 0.0015$ |
| SMOTE | $0.9644 \pm 0.0020$ | $0.0778 \pm 0.0010$ |
| ADASYN | $0.9677 \pm 0.0017$ | $0.0692 \pm 0.0011$ |
| Borderline-SMOTE | $0.9717 \pm 0.0019$ | $0.0679 \pm 0.0013$ |
| CTGAN | $1.0000 \pm 0.0000$ | $0.2796 \pm 0.0227$ |
| TVAE | $1.0000 \pm 0.0000$ | $0.3353 \pm 0.0169$ |

No dataset *Adult*, o Hy-Synergy também alcançou o menor **C2ST = 0.8844** $\pm$

**0.0089** e a menor **JSD = 0.0125** $\pm$ **0.0020** entre os métodos avaliados. Embora o C2ST obtido permaneça distante de $0.5$, ele representa uma redução relevante de separabilidade frente aos demais, uma vez que os métodos generativos profundos atingiram **C2ST = 1.000**, com divergências marginais maiores. Pode-se afirmar que os resultados desse dataset refletem a maior dificuldade de reproduzir com precisão a estrutura conjunta de um domínio misto, mas ainda assim, o ganho relativo frente aos baselines foi consistente.

### 5.3. Utilidade

As metricas de utilidade, também descritas na literatura como métricas de *ML efficiency*, avaliam em que medida os dados sintéticos preservam a capacidade preditiva de modelos treinados com dados reais acrescidos de amostras artificiais (de Boni et al. 2025). Neste estudo, a utilidade foi analisada no protocolo HybridTR, no qual os classificadores são treinados sobre a combinação de dados reais e sintéticos e avaliados em um conjunto de teste exclusivamente real. O desempenho foi quantificado por meio de ROC AUC, que mede a capacidade de discriminação entre classes ao longo de diferentes limiares de decisão, e de F1-Optimal, que resume o equilíbrio entre precisão e revocação no limiar que maximiza esse compromisso.

**Tabela 4. Utilidade preditiva no protocolo HybridTR para o dataset Pima. Valores maiores indicam melhor desempenho.**

| Método | ROC AUC (↑) | F1-Optimal (↑) |
|---|---|---|
| **Hy-Synergy** | $0.8270 \pm 0.0287$ | $0.6689 \pm 0.0381$ |
| CTGAN | $0.5491 \pm 0.1067$ | $0.5181 \pm 0.0061$ |
| SVMSMOTE | $0.5482 \pm 0.1183$ | $0.5174 \pm 0.0054$ |
| SMOTE | $0.5441 \pm 0.1280$ | $0.5147 \pm 0.0126$ |
| ADASYN | $0.5341 \pm 0.1275$ | $0.5150 \pm 0.0107$ |
| Borderline-SMOTE | $0.5334 \pm 0.1241$ | $0.5163 \pm 0.0052$ |
| TVAE | $0.5157 \pm 0.1026$ | $0.5137 \pm 0.0192$ |

Quanto ao dataset Pima (tabela 4), o Hy-Synergy apresentou vantagem sobre todos os métodos comparados, alcançando **ROC AUC =** $0.8270 \pm 0.0287$ e **F1-Optimal =** $0.6689 \pm 0.0381$, enquanto os demais permaneceram entre 0.51 e 0.55 em ROC AUC. Esse contraste indica que a interpolação geométrica e a geração global da distribuição não foram suficientes, nesse domínio, para preservar o sinal preditivo útil ao classificador *downstream*.

**Tabela 5. Utilidade preditiva no protocolo HybridTR para o dataset Adult. Valores maiores indicam melhor desempenho.**

| Método | ROC AUC (↑) | F1-Optimal (↑) |
|---|---|---|
| Hy-Synergy | $0.9122 \pm 0.0111$ | $0.6946 \pm 0.0240$ |
| SVMSMOTE | $0.6413 \pm 0.0386$ | $0.4182 \pm 0.0343$ |
| ADASYN | $0.6392 \pm 0.0395$ | $0.4183 \pm 0.0340$ |
| SMOTE | $0.6391 \pm 0.0398$ | $0.4195 \pm 0.0356$ |
| Borderline-SMOTE | $0.6389 \pm 0.0397$ | $0.4179 \pm 0.0335$ |
| CTGAN | $0.6382 \pm 0.0412$ | $0.4194 \pm 0.0341$ |
| TVAE | $0.6374 \pm 0.0310$ | $0.4180 \pm 0.0229$ |

A tabela 5 mostra que no dataset Adult, o Hy-Synergy atingiu **ROC AUC =** $0.9122 \pm 0.0111$ e **F1-Optimal =** $0.6946 \pm 0.0240$, ao passo que os métodos de referencia permaneceram concentrados em torno de 0.64 de ROC AUC e 0.42 de F1-Optimal.

O ganho indica que a modelagem por *clusters* favoreceu a preservação de padrões uteis para classificação mesmo em um domínio notoriamente mais complexo.

### 5.4. Privacidade

A avaliação de privacidade buscou verificar se a fidelidade e a utilidade observadas poderiam estar associadas à retenção excessiva de informações do conjunto original, ao invés de refletirem um aprendizado distribucional genuíno (de Boni et al. 2025). Para isso, utilizou-se a acurácia de um *Membership Inference Attack* (MIA), cuja interpretação é direta: valores próximos de $0.5$ indicam comportamento próximo ao aleatório e, portanto, menor a evidência de memorização, valores mais altos sugerem maior vulnerabilidade à inferência de pertencimento.

**Tabela 6. Risco de privacidade medido por MIA Accuracy no dataset Pima. O valor ideal é 0.5, correspondente a decisão aleatória.**

| Método | MIA Accuracy (↓) |
|---|---|
| Hy-Synergy | $0.5242 \pm 0.0439$ |
| ADASYN | $0.6610 \pm 0.0473$ |
| SVMSMOTE | $0.6915 \pm 0.0509$ |
| Borderline-SMOTE | $0.6921 \pm 0.0358$ |
| SMOTE | $0.6992 \pm 0.0475$ |
| TVAE | $0.9172 \pm 0.0268$ |
| CTGAN | $0.9190 \pm 0.0328$ |

Os resultados da tabela6 mostram que no dataset Pima, o Hy-Synergy obteve o menor risco relativo, com MIA Accuracy = $0.5242 \pm 0.0439$, próximo do limite de aleatoriedade, enquanto CTGAN e TVAE superaram 0.91 e os métodos geométricos situaram-se entre 0.66 e 0.70. O resultado é consistente com a hipótese de que a síntese guiada por estrutura local reduz a vulnerabilidade à inferência de pertencimento.

**Tabela 7. Risco de privacidade medido por MIA Accuracy no dataset Adult. O valor ideal é 0.5, correspondente a decisão aleatória.**

| Método | MIA Accuracy (↓) |
|---|---|
| Hy-Synergy | $0.6637 \pm 0.0240$ |
| ADASYN | $0.7485 \pm 0.0084$ |
| SVMSMOTE | $0.7492 \pm 0.0120$ |
| Borderline-SMOTE | $0.7535 \pm 0.0070$ |
| SMOTE | $0.7687 \pm 0.0083$ |
| CTGAN | $0.9986 \pm 0.0011$ |
| TVAE | $0.9995 \pm 0.0004$ |

No dataset Adult, os resultados da tabela 7 mostram que o Hy-Synergy apresentou o menor valor entre os geradores comparados **(MIA Accuracy =** $0.6637 \pm 0.0240$**)**, mas ainda distante do ideal de 0.5. Os modelos generativos profundos exibiram acurácias praticamente unitárias. Esses resultados indicam que o método apresentou o melhor compromisso relativo entre os três eixos, mas que há espaço para avanços adicionais na proteção contra memorização em domínios mistos.

### 5.5. Estudo de ablação e análise de estabilidade

Testes de postos sinalizados de *Wilcoxon* confirmam que a remoção da política $\pi(s)$ e do *undersampling* reduz a *ROC-AUC* média em $0,22\%$ e $0,12\%$, respectivamente

($p < 0,05$). O impacto é mais acentuado no *Log Loss*, onde a ausência de *undersampling* causou uma piora de $3,44\%$ (de $0,5116$ para $0,5292$; $p < 0,01$). Crucialmente, a remoção desses componentes disparou o mecanismo *Watchdog* em $4\%$ dos *folds* devido a colapsos de modo, evento inexistente na configuração completa. Em contrapartida, as variantes sem *cleaning* ou *post-processing* não apresentaram diferenças estatísticas significativas ($p > 0,05$). Esses resultados evidenciam que a eficácia do *Hy-Synergy* reside na coordenação entre o diagnóstico topológico e a intervenção seletiva, que atua como um regulador de robustez para o processo generativo.

## 6. Discussão

No dataset *Pima*, o *Hy-Synergy* preservou a estrutura distribucional necessária para sustentar a utilidade sem elevar a vulnerabilidade à inferência de pertencimento (*MIA*). O *C2ST* próximo ao ideal e a menor *JSD* indicam sucesso na captura de correlações em domínios biomédicos contínuos. Já no *Adult*, o contraste entre a baixa *JSD* e o *C2ST* elevado sugere que, em alta cardinalidade, a preservação de distribuições marginais é alcançada mais facilmente que a reconstrução multivariada completa, evidenciando que a fidelidade estrutural absoluta permanece um desafio em dados mistos complexos.

A leitura conjunta dos resultados indica que o ganho do *Hy-Synergy* decorre da sinergia entre diagnóstico topológico e política de decisão local. O estudo de ablação reforça que a remoção da política $\pi(s)$ e do *undersampling* degrada o desempenho, validando a hipótese de que a intervenção regionalizada baseada em *clusters* informativos e zonas de sobreposição é superior à reamostragem uniforme indiscriminada.

As implementações usadas no estágio de geração profunda operaram sob configurações conservadoras. Experimentos em andamento para o refinamento de hiperparâmetros de treinamento (e.g., *batch size* e *learning rate*) sugerem que os resultados aqui apresentados constituem uma estimativa conservadora do potencial do *framework*. A evolução dessas parametrizações tende a elevar o teto de desempenho tanto do *Hy-Synergy* quanto dos *baselines* avaliados.

A aplicação em saúde exige cautela quanto à equidade algorítmica. Embora o *Hy-Synergy* mitigue riscos de reidentificação, ele permanece suscetível à propagação de vieses históricos das bases originais. É imperativo que a síntese de *clusters* não mascare disparidades demográficas em subgrupos minoritários, demandando a integração de auditorias de justiça diagnóstica (*fairness*) para assegurar a aplicabilidade ética em populações heterogêneas.

Do ponto de vista aplicado, esses achados favorecem contextos de saúde com dados escassos e sensíveis. Em bases clínicas restritas, onde a replicação indiscriminada amplifica ruído, a abordagem orientada por estrutura local propicia a construção de modelos mais robustos para apoio ao diagnóstico e estratificação de risco.

## 7. Conclusão

O *Hy-Synergy* demonstrou ser uma abordagem resiliente para a síntese de dados sob escassez e desbalanceamento, superando *baselines* consolidados. Sua contribuição reside na orquestração regionalizada da síntese e na integração sinérgica de técnicas de topologia e diagnóstico local, suprindo a lacuna de avaliações conjuntas de utilidade, fidelidade e privacidade (de Boni et al. 2025).

Embora o método tenha apresentado o melhor compromisso relativo entre as métricas nos domínios avaliados, reconhecem-se limitações quanto à validação em cenários clínicos reais e ao conjunto restrito de *datasets*. A avaliação de privacidade via *MIA*, conquanto relevante, constitui uma caracterização parcial do risco em dados sensíveis.

Trabalhos futuros focarão na expansão da avaliação para bases de maior escala e no refinamento das arquiteturas generativas para mitigar o colapso de modo, além de investigar e desenvolver um módulo específico de avaliação de equidade e a inclusão de protocolos de *Differential Privacy*, garantindo que a geração sintética promova a justiça algorítmica e a proteção rigorosa da privacidade em sistemas de saúde.

## 8. Agradecimentos

This work has been partially funded by the project Startups development supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT of the MCTI grant number 057/2023, signed with EMBRAPII. The authors are also grateful to the Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) for the financial support provided for this research (Grant 64448878/2024).

## Referências

Alshantti, A., Varagnolo, D., Rasheed, A., Rahmati, A., and Westad, F. (2024). Castgan: Cascaded generative adversarial network for realistic tabular data synthesis. *IEEE Access*, 12:13213–13232.

Apellaniz, P. A., Parras, J., and Zazo, S. (2024). An improved tabular data generator with VAE-GMM integration. In *Proceedings of the 32nd European Signal Processing Conference (EUSIPCO)*. IEEE.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

de Boni, M. H. L., Sene Junior, I. G., and Costa, R. M. d. (2025). Tabular data augmentation using artificial intelligence: A systematic review and taxonomic framework. *IEEE Access*, 13.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Eckardt, J.-N., Hahn, W., Röllig, C., Stasik, S., Platzbecker, U., Müller-Tidow, C., Serve, H., Baldus, C. D., Schliemann, C., Schäfer-Eckart, K., Hanoun, M., Kaufmann, M., Burchert, A., Thiede, C., Schetelig, J., Bornhäuser, M., Wolfien, M., and Middeke, J. M. (2023). Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. *Blood*, 142:2268–2268.

Han, H., Wang, W.-Y., and Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing*, volume 3644 of *Lecture Notes in Computer Science*, pages 878–887. Springer.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328. IEEE.

Inan, M. S. K., Hossain, S., and Uddin, M. N. (2023). Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor's morphological information. *Informatics in Medicine Unlocked*, 37.

Kang, H. Y. J., Batbaatar, E., Choi, D. W., Choi, K. S., Ko, M., and Ryu, K. S. (2023). Synthetic tabular data based on generative adversarial networks in health care: Generation and validation using the divide-and-conquer strategy. *JMIR Medical Informatics*, 11.

Napierala, K. and Stefanowski, J. (2016). Types of imbalanced data, differentiation of methods, and appropriate strategies. *Information Sciences*, 330:223–244.

Nguyen, H. M., Cooper, E. W., and Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21.

Panfilo, D., Boudewijn, A., Saccani, S., Coser, A., Svara, B., Chauvenet, C. R., Mami, C. A., and Medvet, E. (2023). A deep learning-based pipeline for the generation of synthetic tabular data. *IEEE Access*, 11:63306–63323.

Patki, N., Wedge, R., and Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics Workshops (DSAAW)*, pages 399–410. IEEE.

Rodriguez-Almeida, A. J., Fabelo, H., Ortega, S., Deniz, A., Balea-Fernandez, F. J., Quevedo, E., Soguero-Ruiz, C., Wagner, A. M., and Callico, G. M. (2023). Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets. *IEEE Journal of Biomedical and Health Informatics*, 27:2670–2680.

Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C., and Johannes, R. S. (1988). Pima indians diabetes database. UCI Machine Learning Repository.

Smith, M. R., Martinez, T., and Giraud-Carrier, C. (2014). An instance level analysis of data complexity. *Machine Learning*, 95(2):225–256.

Wang, W. and Pai, T. W. (2023). Enhancing small tabular clinical trial dataset through hybrid data augmentation: Combining smote and wcgan-gp. *Data*, 8.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32.