Integração de Redes Adversárias Generativas, Transformers e Aprendizado Semi-Supervisionado para Geração de Dados Tabulares Sintéticos

Rafael A. Costa[1] **, Rafael C. Pregardier**[1] **, Isabel C. Reinheimer**[2] **,**
Carlos E. Poli-de-Figueiredo[2] **, Luis A. L. Silva**[1]

[1] Curso de Ciência da Computação e Curso de Sistemas de Informação
Universidade Federal de Santa Maria
Av. Roraima nº 1000 – Santa Maria – RS – Brazil – CEP: 97.105-900

[2]Escola de Medicina (ESMED), Laboratório de Nefrologia
Pontifícia Universidade Católica do Rio Grande do Sul
– Porto Alegre – RS – Brazil – CEP: 90.619-900
{racosta, rcpregardier, luisalvaro}@inf.ufsm.br,
cristinareinheimer@gmail.com, cepolif@pucrs.br

Abstract. *This work proposes the Semi-Supervised Tabular Transformer GAN (STTGAN), an architecture that integrates Generative Adversarial Networks, semi-supervised learning, and Transformer attention mechanisms to generate realistic synthetic tabular data on Chronic Kidney Disease. The approach enables training with few labels while maintaining predictive utility. STTGAN is evaluated using quantitative metrics, with results showing that the architecture can generate data that preserves clinical information even with scarce labels.*

Resumo. *Este trabalho propõe a Semi-Supervised Tabular Transformer GAN (STTGAN), uma arquitetura que integra Redes Adversárias Generativas, aprendizado semi-supervisionado e mecanismos de atenção de Transformers para gerar dados tabulares sintéticos realistas sobre Doença Renal Crônica. A abordagem permite o treinamento com poucos rótulos, mantendo a utilidade preditiva. A STTGAN é avaliada utilizando métricas quantitativas, onde resultados demonstram que a arquitetura permite gerar dados que preservam informações clínicas, mesmo com rótulos escassos.*

1. Introdução

O emprego de técnicas de aprendizado de máquina tem se mostrado promissor no suporte a decisões na área da saúde. Apesar disso, o desempenho de modelos treinados é muitas vezes limitado pela escassez de dados e outras restrições éticas associadas ao uso de registros médicos. Neste contexto, a investigação de arquiteturas de Redes Adversariais Generativas (GANs) [Ghosheh et al. 2024] na geração de dados clínicos sintéticos tem apresentado soluções inovadoras para este problema.

[Liu et al. 2024] revisa modelos de GANs construídos para lidar com dados tabulares mistos (contínuos e categóricos), comuns em problemas na área da saúde. [Schurer et al. 2025] propõe o uso de GANs associadas a técnicas de transferência de aprendizado para gerar dados sintéticos de doenças renais crônicas, importante condição

de saúde também investigada neste trabalho. Tal como no modelo STTGAN proposto, modelos de Transformers [Vaswani et al. 2017] vêm sendo aplicados em dados tabulares por sua capacidade de capturar relações complexas entre atributos. [Kang et al. 2025] integra Transformers ao modelo de GAN, modelando dependências entre variáveis heterogêneas. Investigado no nosso projeto, o Aprendizado Semi-supervisionado (SSL) [Yang et al. 2023] emerge como alternativa para o uso simultâneo de dados rotulados e não rotulados nestes modelos. Ao integrar SSL em GANs, a abordagem resultante visa aumentar a robustez de classificadores, como no modelo SS-GAN [Sricharan et al. 2017], onde o discriminador também aprende a classificar exemplos reais com rótulos escassos.

Este trabalho investiga a combinação de GANs, SSL e mecanismos de atenção de Transformers na geração de dados tabulares para problemas na área da saúde. O trabalho propõe a arquitetura Semi-Supervised Tabular Transformer GAN (STTGAN), a qual reproduz distribuições estatísticas realistas nos dados sintéticos, produzindo dados tabulares úteis para tarefas preditivas, mesmo em contextos com supervisão limitada. Experimentos desenvolvidos focalizam a geração semi-supervisionada de dados para Doenças Renais Crônicas (DRC). DRC é uma condição progressiva que impõe grandes desafios clínicos, exigindo estratégias de diagnóstico precoce e gestão de recursos [Jha et al. 2013].

2. Arquitetura STTGAN

A arquitetura STTGAN é voltada à geração de dados sintéticos clínicos em cenários de escassez de rótulos. A base de dados utilizada é o conjunto CKD-ROUTE [Iimori et al. 2018], que reúne informações de 1.138 pacientes com DRC. A variável-alvo utilizada nos modelos de classificação utilizados neste trabalho é a progressão da doença. Portanto, o trabalho envolve um problema de classificação binária com distribuição de dados desbalanceada, o que é muito comum em várias áreas da saúde.

Para garantir uma modelagem realista, variáveis altamente correlacionadas com o desfecho da doença, como medições futuras da função renal, variáveis de terapia substitutiva e eventos clínicos tardios, foram removidas da base de dados. Variáveis contínuas foram transformadas em categorias clínicas baseadas em faixas de referência, buscando robustez frente a outliers e alinhamento com práticas médicas. Após a seleção de variáveis, os dados foram codificados em valores inteiros, preenchidos com a moda quando ausentes e divididos estratificadamente em conjuntos de treino e teste.

A arquitetura da STTGAN combina mecanismos de atenção multi-cabeça dos Transformers com GANs. O gerador recebe um vetor de ruído concatenado a uma codificação one-hot do rótulo (quando disponível) e o processa por dois blocos Transformer. As saídas incluem amostras categóricas (via Gumbel-Softmax). O discriminador é uma rede profunda que recebe tanto os dados reais quanto os gerados e produz duas saídas: a primeira distingue exemplos reais e sintéticos (real/fake) e a segunda realiza a predição supervisionada do rótulo, quando disponível. A predição supervisionada é treinada apenas com exemplos rotulados, por meio de uma função de perda mascarada, e seus valores são ponderados e somados à perda adversarial para guiar o treinamento do discriminador.

Durante o treinamento, o modelo opera em batches contendo exemplos rotulados e não rotulados em um processo de *self-training* [Yang et al. 2023]. A função de perda do discriminador combina os componentes adversarial e de classificação supervisionada,

ponderados por um hiperparâmetro ajustado empiricamente no nosso projeto. O gerador é otimizado com base na perda adversarial e na entropia das saídas. O treinamento foi conduzido por 80 épocas, com batches de 64 exemplos, vetor latente de dimensão 64, e taxa de aprendizado de 0,0005, com regularização L2. Ao final, os dados gerados são reconstruídos para o formato original por meio de um módulo de transformação reversa, viabilizando sua interpretação e uso em experimentos preditivos com classificadores.

3. Experimentos e Resultados

Os experimentos avaliam a eficácia da STTGAN em diferentes níveis de disponibilidade de rótulos, considerando a utilidade preditiva dos dados sintéticos gerados. Foram investigadas questões relacionadas ao desempenho de modelos preditivos com dados sintéticos, considerando possíveis benefícios da mistura de dados reais e sintéticos.

O conjunto de dados real [Iimori et al. 2018] foi dividido em 80% para treino e 20% para teste, com estratificação das classes. Diferentes cenários de supervisão foram simulados, com percentuais de rótulo variando de 10% a 100%. Para cada cenário, a STTGAN foi treinada utilizando exemplos rotulados (para perda supervisionada) e não rotulados (para o componente adversarial), gerando posteriormente dados sintéticos em quantidade equivalente ao conjunto de treino.

Tabela 1. Desempenho preditivo (F1-Score) utilizando RandomForest e XGBoost para diferentes percentuais de dados rotulados e conjuntos de treino.

% Rótulos	RandomForest			XGBoost		
	Real	Sintético	Semi-Sintético	Real	Sintético	Semi-Sintético
100	0.70	0.72	0.75	0.71	0.71	0.76
90	0.73	0.71	0.78	0.72	0.71	0.76
80	0.80	0.73	0.79	0.74	0.73	0.76
70	0.75	0.73	0.77	0.71	0.72	0.75
60	0.76	0.72	0.77	0.77	0.68	0.76
50	0.73	0.72	0.77	0.76	0.73	0.75
40	0.72	0.71	0.76	0.74	0.68	0.76
30	0.70	0.70	0.74	0.69	0.68	0.70
20	0.68	0.72	0.72	0.71	0.67	0.74
10	0.65	0.60	0.62	0.65	0.54	0.59

A utilidade prática dos dados gerados foi avaliada com classificadores Random Forest e XGBoost, em versões supervisionadas e semi-supervisionadas baseadas no *SelfTrainingClassifier*. A Tabela 1 apresenta os principais resultados preditivos (F1-Score) para os três cenários analisados — real, sintético e semi-sintético — considerando diferentes percentuais de dados rotulados (de 10% a 100%). Observa-se que, em grande parte dos casos, os dados sintéticos e semi-sintéticos atingem F1-scores comparáveis ou superiores ao cenário com 100% de dados rotulados.

Este efeito se mantém especialmente até o limite de 20% de rótulos disponíveis, indicando que os dados sintéticos gerados pela GAN podem complementar e enriquecer o conjunto de treinamento em situações de escassez de rótulos. Os resultados sugerem que a utilização de dados semi-sintéticos oferece uma alternativa viável para aumentar a robustez dos modelos preditivos, mitigando o impacto do desbalanceamento e da limitação de dados rotulados.

A integração de dados reais e sintéticos mostrou-se vantajosa, funcionando como forma de regularização ao enriquecer o conjunto de treino com maior diversidade de padrões. Essa integração beneficiou o desempenho dos modelos RandomForest e XGBoost, sobretudo em cenários com supervisão limitada. A combinação de dados reais e sintéticos superou, em muitos casos, o desempenho obtido com o uso de dados puramente reais, evidenciando o potencial dos dados gerados pela STTGAN.

4. Conclusões

Resultados experimentais evidenciam que a STTGAN proposta é capaz de gerar amostras sintéticas com alta utilidade prática, mesmo sob baixos níveis de supervisão. O emprego de métodos de auto-treinamento permitiu expandir o aprendizado em contextos de rótulos escassos, e os modelos preditivos beneficiaram-se da diversidade trazida pelas amostras geradas. Tais achados são particularmente relevantes para aplicações em saúde, onde o custo de rotulagem é elevado e o volume de dados disponíveis é limitado.

Referências

Ghosheh, G. O., Li, J., and Zhu, T. (2024). A survey of generative adversarial networks for synthesizing structured electronic health records. In *ACM Computing Surveys*, volume 56, pages 1–34.

Iimori, S., Naito, S., Noda, Y., Sato, H., Nomura, N., Sohara, E., Okado, T., Sasaki, S., Uchida, S., and Rai, T. (2018). Data from: Prognosis of chronic kidney disease with normal-range proteinuria: The ckd-route study. Version 1, 2018.

Jha, V., Garcia-Garcia, G., Iseki, K., Li, Z., Naicker, S., Plattner, B., Saran, R., Wang, A. J., and Yang, C. (2013). Chronic kidney disease: global dimension and perspectives. *The Lancet*, 382(9888):260–272.

Kang, H. Y. J., Ko, M., and Ryu, K. S. (2025). Tabular transformer generative adversarial network for heterogeneous distribution in healthcare. *Scientific Reports*, 15(10254).

Liu, T., Fan, J., Li, G., Tang, N., and Du, X. (2024). Tabular data synthesis with generative adversarial networks: design space and optimizations. *The VLDB Journal*, 33:255–280.

Schurer, L., Assunção, J. V. C., Reinheimer, I. C., de Figueiredo, C. E. P., and Silva, L. A. L. (2025). Gans and fine-tuning through transfer learning for the generation of electronic health records on chronic kidney diseases. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE.

Sricharan, K., Bala, R., Shreve, M., Ding, H., Kumar, S., and Sun, J. (2017). Semi-supervised conditional gans. *arXiv preprint arXiv:1708.05789*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Łukasz Kaiser, and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yang, X., Song, Z., King, I., and Xu, Z. (2023). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954.