Além da Similaridade: Uma super-métrica Generalizável para Avaliação de Fidelidade em Dados Sintéticos de Malware

Anna Luiza Gomes da Silva[1], **Angelo Gaspar Diniz Nogueira**[1], **Diego Kreutz**[1]
Kayuã Oleques Paim[2], **Rodrigo Brandão Mansilha**[1], **Celso Nobre da Fonseca**[1]

[1] Horizon IA Labs e PPGES, Universidade Federal do Pampa (UNIPAMPA)

[2]INF, Universidade Federal do Rio Grande do Sul (UFRGS)

angelonogueira.aluno@unipampa.edu.br, kopaim@inf.ufrgs.br
diegokreutz@unipampa.edu.br, mansilha@unipampa.edu.br
annaluiza.aluno@unipampa.edu.br, celsofonseca@unipampa.edu.br

Resumo. *Este trabalho propõe uma super-métrica flexível para a avaliação de dados sintéticos de malware, integrando oito medidas-chave (como RMSE, Jaccard e Wasserstein) em quatro dimensões principais: Distância, Correlação/Associação, Similaridade de Características e Distribuição Multivariada. A super-métrica permite o ajuste dinâmico de pesos entre essas dimensões, adaptando-se a diferentes cenários e objetivos de análise. A proposta não tem como finalidade substituir métricas existentes, mas sim oferecer um framework integrado e adaptável para avaliação multidimensional. Validada em quatro datasets de malware Android, utilizando o modelo TVAE, a super-métrica demonstrou forte correlação com a métrica de utilidade recall, além de apresentar estabilidade estatística significativamente superior à Similaridade do Cosseno, com desvio padrão três vezes menor (0,026 vs. 0,083) e menor amplitude (0,06 vs. 0,20) entre os conjuntos avaliados. Os resultados comprovam a eficácia da super-métrica na avaliação simultânea de aspectos locais e globais dos dados sintéticos, oferecendo maior robustez estatística em comparação com métricas convencionais.*

1. Introdução

A geração de dados sintéticos em cibersegurança, fundamentada em modelos de linguagem, é uma necessidade cada vez mais comum pois permite superar a escassez de dados reais e aprimorar a detecção de ameaças. Estudos recentes demonstram que a combinação de métricas de fidelidade e utilidade (acurácia em classificação) valida a eficácia desses dados [Almorjan et al. 2025].

A avaliação da qualidade de dados sintéticos representa, contudo, um desafio significativo, evidenciado pela ampla variedade de métricas desenvolvidas para esse fim. Uma revisão sistemática da literatura identificou 65 métricas distintas utilizadas para essa finalidade. Longe de representar uma solução consolidada, essa diversidade tem provocado uma fragmentação metodológica [Ibrahim et al. 2025, Pezoulas et al. 2024, Murtaza et al. 2023], dificultando a padronização e a comparação entre estudos.

Cada pesquisa tende a adotar um subconjunto distinto de métricas. Por exemplo, [Sun et al. 2023] utiliza distância de Hamming, distância Euclidiana e erro de regressão linear como indicadores de fidelidade, enquanto [Xin et al. 2020] emprega a distância de

Wasserstein e a *Fréchet Inception Distance* (FID). Essa heterogeneidade compromete a comparabilidade sistemática entre diferentes modelos e conjuntos de dados.

O problema central, portanto, não reside na ausência de métricas, mas na falta de uma abordagem capaz de sintetizar, de forma flexível e adaptável, as múltiplas facetas da fidelidade de dados sintéticos. As métricas atualmente disponíveis são, em geral, rígidas e focadas em aspectos isolados, não permitindo que os pesquisadores ajustem a avaliação conforme as especificidades de seus cenários. Em determinados contextos, por exemplo, a preservação das correlações entre *features* pode ser mais relevante do que a aderência à distribuição estatística global, mas as abordagens existentes não oferecem mecanismos simples para ponderar essa prioridade.

Para superar essa limitação, este trabalho introduz o conceito de uma "super-métrica" maleável, capaz de agregar os sinais de diversas métricas de base, permitindo o ajuste de pesos de acordo com as prioridades definidas para cada cenário e contexto específico de avaliação. Em vez de propor mais uma métrica isolada, apresenta-se um *framework* que sintetiza informações provenientes de métricas já consolidadas, oferecendo uma visão holística, personalizável e mais informativa sobre a qualidade dos dados sintéticos.

A ausência de consenso quanto às métricas e valores de referência não apenas dificulta a quantificação sistemática da fidelidade dos dados sintéticos e limita a comparabilidade entre estudos, mas, de forma ainda mais crítica, impede que pesquisadores obtenham feedback consistente e confiável para o ajuste de hiperparâmetros e arquiteturas de modelos generativos. Abordagens recentes destacam a urgência de métricas mais universais, capazes de superar a fragmentação atual e promover avaliações mais robustas [Chundawat et al. 2022].

Além disso, um desafio crucial frequentemente negligenciado pelas métricas de fidelidade é a plausibilidade dos dados gerados. Como formalizado por [Cortellazzi et al. 2024], há uma lacuna fundamental entre o espaço do problema (isto é, o mundo real, por exemplo, um *malware* funcional) e o espaço de características (a representação vetorial manipulada pelo modelo). A geração de dados sintéticos que otimiza exclusivamente a similaridade no espaço de características pode produzir amostras estatisticamente fiéis, mas inviáveis ou sem sentido no espaço do problema — como, por exemplo, um *malware* com combinações de permissões mutuamente exclusivas. Tais dados, embora válidos do ponto de vista métrico, tornam-se inúteis para o treinamento e a avaliação de modelos aplicados ao mundo real.

Para a avaliação inicial da super-métrica proposta, utilizamos quatro *datasets* de *malware* Android (Adroit, Androcrawl, Drebin215 e Kronodroid Emulator), em conjunto com o modelo *Table Variational Autoencoder* (TVAE) da SDV [Xu et al. 2019].

2. Proposta da super-métrica

A super-métrica não pretende invalidar métricas individuais, mas integrar suas contribuições em um escore único e ajustável. Essa abordagem permite refletir a realidade multifacetada da fidelidade, onde diferentes cenários podem demandar ênfases distintas (e.g., distâncias em dados binários vs. distribuições multivariadas em dados contínuos).

Nesta seção, apresentamos a metodologia e o resultado da elaboração da

super-métrica. A Figura 1 apresenta uma visão geral do processo. Primeiramente conduziu-se uma revisão sistemática da literatura seguindo as diretrizes metodológicas de [Kitchenham et al. 2009].

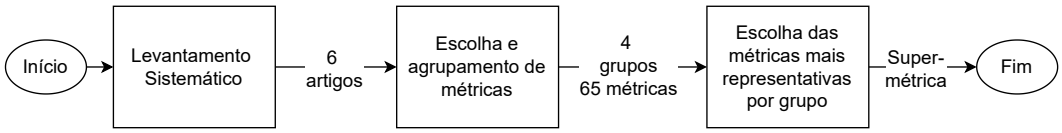

Figura 1. Fluxograma sobre a metodologia.

Como resultado foram identificados seis trabalhos ([Ibrahim et al. 2025, Pezoulas et al. 2024, Murtaza et al. 2023, Perkonoja et al. 2024, Boudewijn et al. 2023, Hernadez et al. 2023]), que permitiram o mapeamento de 65 métricas de fidelidade. Posteriormente elas foram categorizadas em 4 grupos, e o agrupamento das métricas foi realizado usando dois métodos complementares sendo elas: (i) a categorização original proposta pelos estudos; (ii) análise realizada pelos autores.

A seleção das métricas por grupo considerou a frequência de citações na literatura, buscando utilizar as métricas mais usadas, na Tabela 1 é apresentada o conjunto de métricas selecionadas, incluindo sua classificação por grupos sendo eles a, b, c e d e as respectivas siglas.

Tabela 1. Grupos e métricas representativas escolhidas.

Grupo	Sigla	Métrica	#
Métricas de Distância (a)	RMSE	RMSE (erro quadrático médio)	1
	HAD	Distância de Hamming	2
	CS	Similaridade de cosseno	3
Correlação e Associação (b)	JSI	Índice de Similaridade de Jaccard	4
Similaridade de características (c)	FCO	Co-ocorrência de características	5
	FDS	Similaridade de Distribuição de Características	6
Distribuição Multivariada (d)	WD	Distância de Wasserstein	7
	HED	Distância de Hellinger	8

De acordo com as informações da Tabela 1 discute-se em sequência os grupos de métricas, sendo elas:

- **Métricas de Distância (Grupo A)**. Este grupo avalia a "proximidade" fundamental entre os dados reais e sintéticos. A seleção de RMSE, Distância de Hamming (HD) e Similaridade de Cosseno (CS) foi intencional para capturar dimensões complementares de fidelidade. O RMSE quantifica a magnitude do erro em valores contínuos, indicando se os dados sintéticos mantêm escalas realistas. A Distância de Hamming é crucial para dados de *malware* com muitas características binárias (e.g., permissões), medindo o número exato de discrepâncias. Já a Similaridade de Cosseno foca na preservação dos padrões direcionais e da orientação dos vetores de características, independentemente da magnitude. Usá-las em conjunto evita que uma única métrica mascare falhas: por exemplo, dados com alta similaridade de cosseno (mesma direção) poderiam ter erros de magnitude (RMSE alto) ou discrepâncias binárias (HD alto) que passariam despercebidos.
- **Correlação e Associação (Grupo B)**. Este grupo foca na preservação das relações entre características. O Índice de Jaccard (JSI) foi escolhido por ser especialmente

eficaz para dados binários, medindo a coocorrência de características, um aspecto fundamental para a modelagem de dependências em dados de *malware*.

- **Similaridade de Características (Grupo C)**. Enquanto a correlação mede a relação entre pares, este grupo avalia a fidelidade das distribuições univariadas e bivariadas. Métricas como Co-ocorrência de Características (FCO) e Similaridade de Distribuição de Características (FDS) garantem que as propriedades estatísticas de características individuais ou em pares sejam mantidas, evitando que o modelo generativo distorça padrões fundamentais dos dados.
- **Distribuição Multivariada (Grupo D)**. Este é o grupo mais global, avaliando a similaridade da distribuição conjunta de todas as características. A Distância de Wasserstein (WD) e a Distância de Hellinger (HED) são métricas robustas que capturam diferenças complexas entre distribuições de probabilidade, superando as limitações de métricas que avaliam apenas momentos estatísticos (como média e variância).

A super-métrica proposta estabelece um escore unificado de fidelidade através do seguinte fluxo: (i) normalização individual de cada métrica para o intervalo [0,1]; (ii) agregação intra-grupo com pesos unitários; e (iii) combinação intergrupos ponderada conforme a equação:

$$SM = \frac{w_1\mathcal{A} + w_2\mathcal{B} + w_3\mathcal{C} + w_4\mathcal{D}}{w_1 + w_2 + w_3 + w_4}$$

onde $\mathcal{A}$ representa as métricas de Distância, $\mathcal{B}$ de Correlação, $\mathcal{C}$ de Similaridade, e $\mathcal{D}$ de Distribuição Multivariada, E $w_1 = 3$, $w_2 = 3$, $w_3 = 1$, $w_4 = 1$ são os coeficientes de peso ajustáveis. Esses coeficientes foram ajustados empiricamente com um rationale claro que visou priorizar os aspectos de fidelidade mais fundamentais e que demonstraram maior impacto na utilidade prática dos dados. Para isso utilizou-se o alinhamento com a métrica de utilidade *recall* como guia principal para a calibração; onde observou-se que os grupos de Distância ($\mathcal{A}$) e Correlação ($\mathcal{B}$), que capturam a fidelidade em nível de característica e de relacionamento, eram os mais sensíveis para refletir a qualidade dos dados em tarefas de classificação, e nesse caso receberam um peso maior. Os grupos de Similaridade de Características ($\mathcal{C}$) e Distribuição Multivariada ($\mathcal{D}$), embora importantes para a fidelidade global, receberam um peso menor para equilibrar a avaliação, evitando que um bom desempenho em estatísticas globais mascarasse algumas falhas que seriam evidenciadas em outros grupos. Essa ponderação de valores visa, então, criar uma métrica que seja estatisticamente fiel, sendo a estabilidade resultante dessa agregação ponderada uma característica de design intencional e desejável.

Ao combinar métricas complementares que avaliam diferentes dimensões da fidelidade, a super-métrica é projetada para ser robusta a flutuações extremas em um único aspecto da avaliação. Essa compensação mútua não visa mascarar falhas, mas sim fornecer uma avaliação mais equilibrada e consistente da qualidade geral dos dados, o que é essencial para comparações sistemáticas e confiáveis entre diferentes modelos generativos.

[da Fonseca 2024] adotou uma abordagem semelhante voltada à avaliação de critérios de informação, com a formulação de um critério adaptável baseado nos coeficientes de superfícies quádricas. Os resultados desse trabalho indicaram que o uso

de penalidades variáveis permitiu maior flexibilidade na seleção de modelos, especialmente em contextos com séries temporais geradas por diferentes estruturas autoregressivas e de médias móveis, Nessa formulação, o autor observou que, em determinadas situações, o critério proposto apresentou desempenho compatível e, em alguns casos, ligeiramente superior quando comparado aos critérios tradicionais, como AIC, AICc e BIC. Seguindo caminho semelhante, os coeficientes da super-métrica foram ajustados com o objetivo de refletir a utilidade prática dos dados sintéticos (por meio da comparação com o *recall*), buscando superar as limitações das métricas convencionais que, como alertam [Figueira and Vaz 2022], podem apresentar estatísticas descritivas semelhantes enquanto mascaram diferenças distributivas substanciais.

3. Avaliação

3.1. Datasets

Os quatro datasets selecionados para este estudo são amplamente utilizados na literatura sobre detecção de *malware* [Nawshin et al. 2024, Alomari et al. 2023, Wang et al. 2019], sendo suas especificações detalhadas na Tabela 2. Uma cópia desses conjuntos de dados foi obtida a partir do repositório do projeto MalwareDataHunter[1], cujo um dos objetivos foi catalogar conjuntos de dados tipicamente utilizados no domínio de detecção de *malware*. A tabela detalha o número de amostras, características e tipo de características. De acordo com a documentação do repositório foram selecionadas as 200 características mais importantes com base no método *chi-square*, e cada *dataset* foi balanceado para um total máximo de 10.000 amostras por classe.

Tabela 2. Datasets considerados neste estudo

Dataset	Amostras	Características	Tipo de Características
ADROIT	6.836	118	Permission
AndroCrawl	20.340	136	Permission
Drebin215	11.110	200	Permissions, Intents, API Calls
KronoDroid Emulator	20.000	200	Permissions, Intents, API Calls

3.2. TVAE

O TVAE (*Temporal Variational Autoencoder*) é uma arquitetura neural baseada em *variational autoencoder* (VAE) [Kingma et al. 2013], especificamente desenvolvida para geração de dados sintéticos tabulares pelo projeto SDV[2]. Em nossos experimentos, os dados sintéticos produzidos por esta rede em processo de avaliação cruzada de 5 dobras (*folds*), são avaliados através do classificador KNN (*K-nearest neighbors*), para obter a métrica de *recall*, que foi adotada como métrica de utilidade para os experimentos feitos.

3.3. Resultados

A Figura 2 representa os resultados obtidos com o modelo TVAE, mensurando a supermétrica, o *recall* com o classificador KNN e o cosseno de similaridade, para cada um dos *datasets* considerados. As métricas representam a média obtida entre uma avaliação cruzada de 5 dobras.

[1] `https://github.com/MalwareDataLab/Datasets/tree/44b14d78f1361a2300daa42b3d4127df8fad7068/JBCS_2025`

[2] `https://sdv.dev`

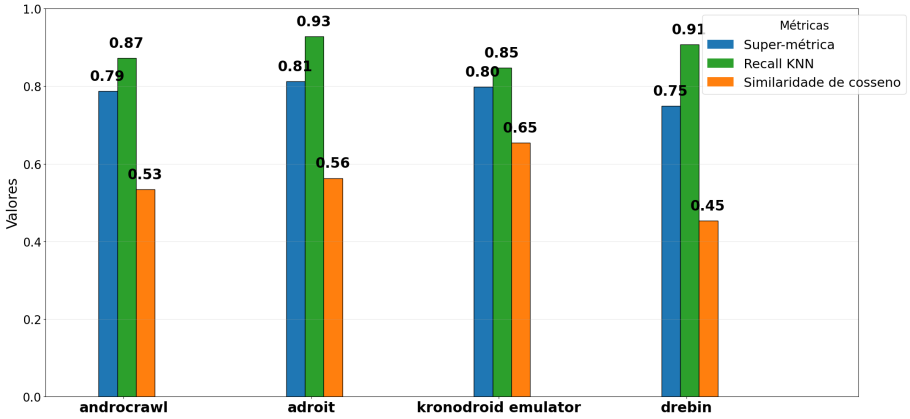

Figura 2. Métricas resultantes para os 4 datasets utilizando o modelo TVAE

A Figura 2 revela duas observações fundamentais: (i) a super-métrica demonstra alinhamento consistente com a tendência do *recall* do classificador knn, apresentando valores comparáveis ou superiores ao *Recall* KNN na maioria dos *datasets*. Sendo generalizada para quatro *datasets* de *malware* com o mesmo comportamento.

Com base nestas evidências, a partir dos dados da Figura 2 postulamos como hipótese que a super-métrica desenvolvida estabelece um paradigma híbrido de avaliação, sendo eles: (1) o alinhamento comportamental com métricas de classificação (como o *recall*), que quantificam a utilidade dos dados sintéticos em tarefas de classificação; e (2) estabilidade superior à similaridade de cosseno. Essa afirmação é evidenciada por menor variabilidade encontrada pela super-métrica e o valor da similaridade de cosseno, com valor de desvio padrão (0,026 vs. 0,083) respectivamente e de amplitude (0,06 vs. 0,20) respectivamente entre *datasets*. Este resultado de estabilidade não é um mero artefato estatístico, mas a validação empírica do nosso design. A suavização observada, onde variações negativas em uma métrica podem ser compensadas por outras, demonstra a robustez da super-métrica em fornecer um escore consistente através de diferentes *datasets*. Crucialmente, como mostra o alinhamento com o *Recall* (Figura 2), essa estabilidade não compromete a sensibilidade da métrica para detectar a utilidade prática dos dados. Isso confirma que a super-métrica atinge o equilíbrio desejado: é estável o suficiente para comparações padronizadas, mas sensível o suficiente para refletir a qualidade real dos dados sintéticos. Esses índices em conjunto com a proximidade ao *recall* é particularmente relevante, uma vez que o *recall* quantifica a proporção de amostras classificadas corretamente como *malware* ou benignas.

Essa capacidade de calibração e o alinhamento com a utilidade posicionam a super-métrica como um candidato promissor para guiar o refinamento e a otimização de modelos generativos, fornecendo um sinal de desempenho mais informativo, similar ao que [Chundawat et al. 2022] observaram com o TabSynDex ao monitorar o aprendizado em diferentes épocas.

4. Limitações e Trabalhos Futuros

Embora a super-métrica proposta avance na avaliação de fidelidade, reconhece-se que sua formulação atual opera primariamente no espaço de características. Isso não garante a plausibilidade dos dados no espaço do problema [Cortellazzi et al. 2024]. Amostras

sintéticas podem ser estatisticamente semelhantes, mas conter combinações de características que são funcionalmente inviáveis no domínio de *malware* Android. Para abordar a questão da plausibilidade, em trabalhos futuros se concentrará em estender a super-métrica com restrições específicas do espaço do problema, incluindo: (i) uma análise de viabilidade de combinações de *features* baseada no conhecimento de domínio (e.g., identificando permissões Android mutuamente exclusivas); e (ii) a incorporação de detectores de anomalias estruturais para validar que os dados sintéticos não sejam apenas estatisticamente fiéis, mas também praticamente viáveis.

Como perspectivas futuras, destacamos: (i) uma análise macro e micro dos pesos das métricas que compõem a super-métrica, incluindo a inclusão dinâmica de novas métricas e a autocalibração de pesos com base em domínios específicos; (ii) a utilização de um maior número de *datasets* e a inclusão de mais classificadores além do KNN; (iii) a realização de testes da super-métrica em casos de colapso modal para comprovar a hipótese levantada; e (iv) a comparação do comportamento da super-métrica em cenários além do Android, explorando a definição dinâmica de pesos entre métricas ou grupos para diferentes tipos de dados.

5. Conclusão

Esta pesquisa propõe uma super-métrica adaptável ao contexto, capaz de integrar múltiplas dimensões de fidelidade e utilidade. Os experimentos demonstraram alinhamento com o *recall* e desempenho superior a métricas isoladas, como a Similaridade do Cosseno, ao capturar aspectos locais e globais dos dados. Validada frente a critérios consolidados, a super-métrica mostrou-se eficaz na avaliação da qualidade dos dados sintéticos, com potencial para mitigar inflacionamentos métricos e oferecer análises mais robustas e contextualizadas.

Agradecimentos. A pesquisa contou com apoio da RNP (Programa Hackers do Bem - GT Malware DataLab), da CAPES (Código de Financiamento 001), da FAPERGS, por meio do edital 02/2022 e dos termos de outorga 24/2551-0001368-7 e 24/2551-0000726-1, e da FAPESP (processos 2020/05183-0 e 2023/00816-2).

Referências

Almorjan, A., Basheri, M., and Almasre, M. (2025). Large language models for synthetic dataset generation of cybersecurity indicators of compromise. *Sensors*, 25(9).

Alomari, E. S., Nuiaa, R. R., Alyasseri, Z. A. A., Mohammed, H. J., Sani, N. S., Esa, M. I., and Musawi, B. A. (2023). Malware detection using deep learning and correlation-based feature selection. *Symmetry*, 15(1):123.

Boudewijn, A., Ferraris, A. F., Panfilo, D., Cocca, V., Zinutti, S., Schepper, K. D., and Chauvenet, C. R. (2023). Privacy measurement in tabular synthetic data: State of the art and future research directions.

Chundawat, V., Tarun, A., Mandal, M., Lahoti, M., and Narang, P. (2022). A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence*, PP:1–11.

Cortellazzi, J., Pendlebury, F., Arp, D., Quiring, E., Pierazzi, F., and Cavallaro, L. (2024). Intriguing properties of adversarial ml attacks in the problem space [extended version].

da Fonseca, C. N. (2024). *Uma abordagem metodológica para a construção de critérios de informação a partir de superfícies quádricas*. Tese de doutorado, FURG.

Figueira, A. and Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and gans. *Mathematics*, 10(15).

Hernadez, M., Epelde, G., Alberdi, A., Cilla, R., and Rankin, D. (2023). Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf Med*, 62(S 01):e19–e38.

Ibrahim, M., Khalil, Y. A., Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G., and Dumontier, M. (2025). Generative ai for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, 189:109834.

Kingma, D. P., Welling, M., et al. (2013). Auto-encoding variational bayes.

Kitchenham, B., Brereton, P., Budgen, D., Turner, M., Bailey, M., and Linkman, S. (2009). Systematic literature reviews in software engineering – a systematic literature review. *Information and Software Technology*, 51(1):7–15.

Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., and Bano, A. (2023). Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546.

Nawshin, F., Gad, R., Unal, D., Al-Ali, A. K., and Suganthan, P. N. (2024). Malware detection for mobile computing using secure and privacy-preserving machine learning approaches: A comprehensive survey. *Computers and Electrical Engineering*, 117:109233.

Perkonoja, K., Auranen, K., and Virta, J. (2024). Methods for generating and evaluating synthetic longitudinal patient data: a systematic review.

Pezoulas, V. C., Zaridis, D. I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N. S., and Fotiadis, D. I. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. *CSBJ*, 23:2892–2910.

Sun, C., van Soest, J., and Dumontier, M. (2023). Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics*, 143:104404.

Wang, W., Zhao, M., Gao, Z., Xu, G., Xian, H., Li, Y., and Zhang, X. (2019). Constructing features for detecting android malicious applications: issues, taxonomy and directions. *IEEE access*, 7:67602–67631.

Xin, B., Yang, W., Geng, Y., Chen, S., Wang, S., and Huang, L. (2020). Private fl-gan: Differential privacy synthetic data generation based on federated learning. In *Icassp 2020-2020 IEEE ICASSP*, pages 2927–2931. IEEE.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Adv Neural Inf Process Syst*.